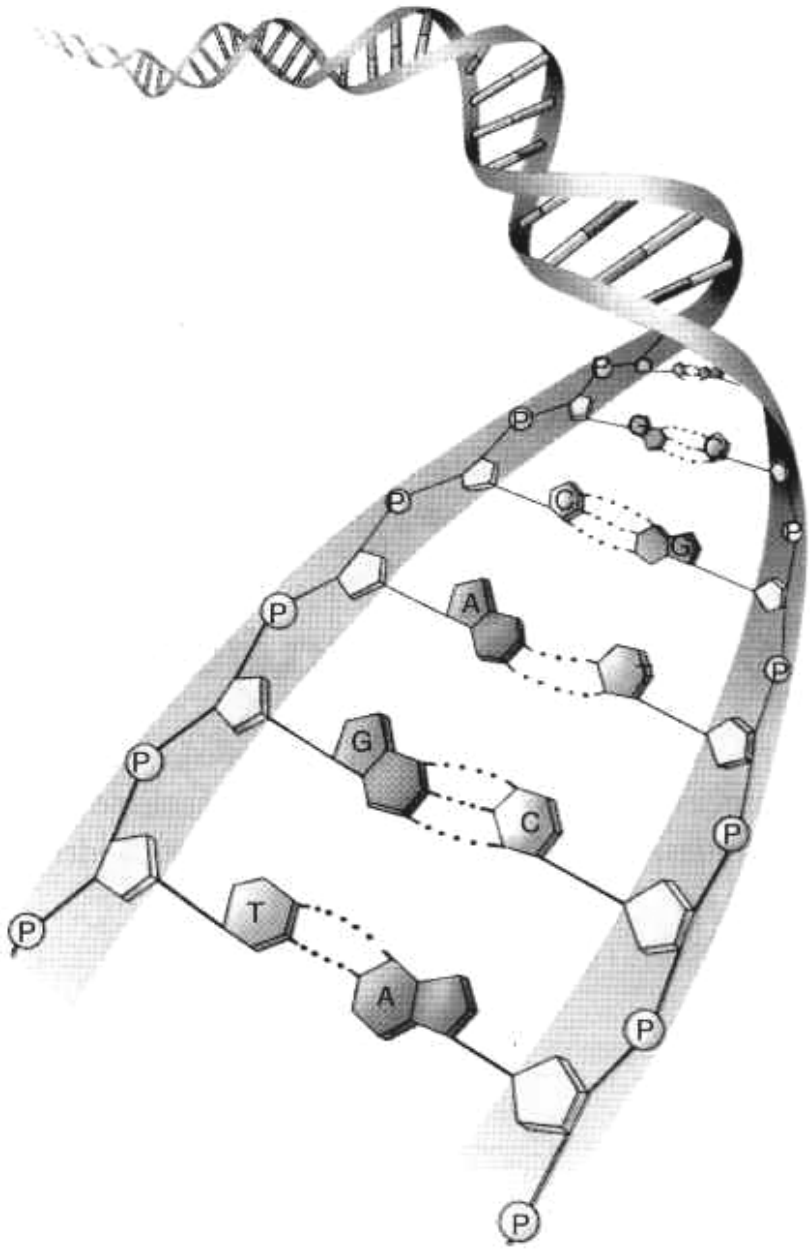
Universidade de Aveiro Departamento de Matemática
2007

Carla Susana Fontinha Vieira

Estudo de Variáveis Discretas: um contributo ao Ensino e à Genética



Universidade de Aveiro Departamento de Matemática
2007

Carla Susana Fontinha Vieira

Estudo de Variáveis Discretas: um contributo ao Ensino e à Genética

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática (perfil de Ensino), realizada sob a orientação científica da Prof.ª Doutora Adelaide de Fátima Baptista Valente Freitas, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro

agradecimentos

À minha orientadora, Prof.ª Doutora Adelaide Freitas, pela sua orientação científica, apoio e disponibilidade.

A todos os colegas de Mestrado, pelo espírito de equipa que se instalou no grupo o que proporcionou o ultrapassar de muitos obstáculos.

À minha família, que me apoia nas etapas da minha vida.

A ti simplesmente: AOE.

o júri

presidente

Professora Doutora Maria Paula Macedo Rocha Malonek
Professora Catedrática do Departamento de Matemática da Universidade de Aveiro

vogais

Prof.ª Doutora Luzia Augusta Pires Gonçalves
Professora Auxiliar da Unidade de Epidemiologia e Bioestatística do Instituto de Higiene e Medicina Tropical da Universidade Nova de Lisboa

Prof.ª Doutora Adelaide de Fátima Baptista Valente Freitas
Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro

pensamento

“A mente que se abre a uma nova ideia jamais volta ao seu tamanho original.”

Albert Einstein

palavras-chave

Binomial, Multinomial, Poisson, testes simultâneos, teoremas limite, DNA, mRNA, codão.

resumo

Nos últimos anos tem-se vindo a assistir a um avanço extraordinário da Genética com a descodificação do código genético completo de um número cada vez maior de espécies.

Usando uma terminologia matemática, pode-se dizer que a molécula de DNA, de qualquer organismo, é constituída por pares de sequências complementares de símbolos retirados de um alfabeto de quatro símbolos (os nucleótidos). Para cada sequência, e dependendo do contexto de interesse, podem ser realizadas contagens daqueles símbolos e daí extrair informações estatísticas relevantes na identificação de regras que governam as sequências de DNA e, consequentemente, a tradução do mRNA pelo ribossoma.

O grupo de Bioinformática da Universidade de Aveiro tem vindo a investigar o sequenciamento de codões (tripletos de nucleótidos) em zonas codificantes do DNA de organismos de qualquer domínio da Vida: *Archaea*, *Eukarya* e *Bacteria*. Um dos objectivos práticos da presente dissertação é contribuir para o avanço dessa investigação analisando propriedades associadas aos pares de codões iguais e aos codões raros. Para tal, cento e dezanove espécies dos três domínios da Vida são aqui consideradas.

Nesta dissertação estudam-se, assim, metodologias estatísticas apropriadas à análise de dados discretos de populações Binomiais, Multinomiais e de Poisson, sendo consideradas diferentes propostas de inferir os parâmetros das distribuições.

Analisam-se e comparam-se ainda, vários procedimentos que garantem a significância global fixada na realização de testes simultâneos. Este estudo permitiu estabelecer diferentes formas de identificar pares de codões problemáticos no sequenciamento do DNA em zonas codificantes.

De forma ainda a contemplar a componente Ensino, foram criadas aplicações interactivas com o propósito de dinamizar o interesse pelo estudo do Tema: Probabilidades e Análise Combinatória, do actual programa do 12º ano de escolaridade. Nessas aplicações mostram-se representações gráficas de funções de massa de probabilidade das distribuições Binomial, Hipergeométrica e de Poisson e tem como objectivo que o aluno/utilizador, de modo intuitivo, estabeleça distribuições limite para as distribuições discretas; nomeadamente, a convergência da distribuição Binomial à Normal, da Hipergeométrica à Binomial e desta à distribuição de Poisson.

keywords

Binomial, Multinomial, Poisson, multiple tests, limit theorems, DNA, mRNA, codon.

abstract

In the past few years we have been assisting a remarkable development of Genetic with the decoding of the genetic code for a larger number of species.

Using a mathematic terminology one can state that the DNA molecule is defined by pairs of complementary sequences of four symbols (nucleotides). Several estimating can be made for these sequences depending on the objective of the studies. It can provide new insight on the rules that govern the DNA sequences and so the mRNA decoding accuracy by the ribosome.

The Aveiro University Bioinformatics group has been investigating the evolution of codons (nucleotides triplets) context on a genome wide scale for species of any Life domain: *Archaea*, *Eukarya* and *Bacteria*. One of the main practical aims of this work is to contribute for the progress of these investigations, by the analysis of associated properties to equal pairs of codons and rare codons. For that matter a hundred and nineteen species from all the three domains were considerer in this study.

In this work, different specifics statistical methodologies to discrete data from Binomial, Multinomial and Poisson populations were studied, by analyzing several approaches to infer their parameters.

Also, several procedures were analysed and compared, that guarantee the global significance of performing multiple tests. These studies allowed us to establish different forms in the identification of problematic codon pairs in coding sequences of genome.

The educational component was also considered in this work by the creation of interactive applications in contents of Mathematic discipline in Secondary school. These applications show graphical representations of the probability mass function for discrete probability distributions herein studied and provide an easy way for students/users to analyse their limiting distributions.

Conteúdo

Capítulo 1

Introdução

1.1 Enquadramento

Fazendo um balanço da investigação científica do século XX, é possível inferir que a primeira metade foi dominada pelas Ciências Físicas, enquanto que a Biologia dominou a segunda. Na realidade, os avanços das Ciências Físicas, e o consequente desenvolvimento de instrumentos e tecnologias (microscópio, radioisótopos, computador, entre outros) permitiram uma posterior focalização em torno da Biologia e da Medicina, no final do século.

Segundo diversos autores, as estruturas vivas são altamente complexas, sendo o cerne do problema o próprio ácido desoxirribonucleico (DNA - DeoxyriboNucleic Acid): esta longa molécula, é a chave do enigma da vida. Muitos consideram que nela se encerra o aspecto exterior de cada espécie biológica, a duração da sua vida e os limites do seu potencial. Esta molécula determina, também, ao mínimo pormenor o que pode fazer cada planta ou animal e claro, o que pode fazer cada célula duma planta ou dum animal. O DNA é considerado o fio da vida, e desde 1950 a explicação progressiva da sua estrutura e funções têm sido uma das preocupações centrais da Biologia (especificamente da Biologia Molecular e da Bioquímica). Tal comportou um grande desenvolvimento na investigação na área da Engenharia Genética. A uma velocidade exponencial um elevado número de sequências de DNA têm ficado disponíveis, tendo o mundo da Genética ficado mergulhado em enormes conjuntos de dados desafiando investigadores da Estatística, Informática e Biologia a formarem equipas de investigação com o objectivo de extrair informação e propriedades estatísticas relevantes contidas no DNA.

Na Universidade de Aveiro existe um grupo interdisciplinar que envolve matemáticos, engenheiros informáticos, bioquímicos e biólogos com o objectivo de contribuir para responder a

conjecturas que se vão formulando dentro desta temática. O tema da presente dissertação surge no seguimento dos trabalhos desenvolvidos no âmbito dos projectos *New bioinformatics tool for genome analysis unvails new rules governing speed and accuracy of mRNA decoding* e PTDC/Mat/72974/2006: "Novas Metodologias estatísticas para análise de dados de microarrays de DNA", ambos financiados pela Fundação para a Ciência e Tecnologia (FCT).
Com a presente dissertação pretende-se dar um contributo à identificação de regras de descodificação do código genético, através da aplicação de metodologias estatísticas apropriadas a dados discretos até agora não consideradas nos referidos projectos.
Além disso, uma vez que a dissertação é no âmbito de um mestrado com perfil direccionado para o Ensino, apresenta-se também uma proposta de abordagem, simples e interactiva, de algumas distribuições discretas e relações limite entre elas, inserindo-a nos conteúdos programáticos da disciplina de Matemática A do 12.º ano.

1.2 Contextualização Biológica

Num primeiro passo, foi importante dominar alguns conceitos básicos da Biologia Molecular e da Genética ([53]).

A Terra cobriu-se de vida, com formas e habitats muito diversificados.
Há borboletas inofensivas com camuflagem que lembra vespas,
e vespas que se assemelham a formigas.
Há aves que ficam no ar, batendo as asas à procura de néctar.
Há tartarugas que nadam milhares de quilómetros,
para deixarem os ovos em determinadas regiões.
Quem não conhece a maravilhosa cauda do pavão?
E o grande pescoço de uma girafa?
E o corpo listado de uma zebra?
As águias têm uma visão maravilhosa!
Já as toupeiras, governam-se bem no interior da terra sem nada verem.
Há também plantas...
Como as sequóias, que podem atingir mais de uma centena de metros de altura.
Outras, como a "sensitiva" têm pequenos folíolos que podem dobrar-se, transformando as frondes verdes em raminhos aparentemente nus.
E o variado mundo dos microrganismos?

Já se identificaram cerca de dois milhões de espécies
mas a total diversidade da vida está estimada em 30 a 40 milhões!
Mas afinal porque diferem os seres vivos uns dos outros?
(in "Terra Universo de Vida" [53])

Todos os organismos existentes na Terra estão divididos em três grandes domínios: Eukarya, Bacteria e Archaea, conforme Figura 1.1. No primeiro grupo encontram-se todos os animais, plantas e fungos; no segundo as bactérias; e no terceiro, seres estranhos como os extremófilos (que vivem em ácido sulfúrico com um pH próximo de zero). Este último domínio só foi identificado há cerca de 25 anos, e ainda se reveste de muito mistério.

Uma explicação para a diversidade da vida reside no DNA pois cada organismo contém material genético, sobre a forma de DNA, o qual contém a informação biológica que define as suas características e controla todas as suas actividades vitais.

Os organismos estão divididos em procariontes, que engloba os Archaea e Bacteria, e eucariontes, os Eukarya.

Nos seres procariontes o DNA encontra-se livre no citoplasma, não estando rodeado por invólucro nuclear. Em cada célula existe normalmente uma única molécula de DNA, de forma circular, que só transitoriamente está associada a proteínas.

Nas células eucariontes o DNA localiza-se no núcleo, estando cada uma destas moléculas associadas a proteínas. O complexo DNA-proteínas é designado por cromatina e o número de filamentos de cromatina em cada célula é variável de espécie para espécie, sendo o termo cromossoma utilizado para designar cada unidade morfológica e fisiológica de cromatina.

Historicamente o DNA foi ignorado pelos biólogos durante quase um século após a sua descoberta, em 1868, por F. Miesher. Até 1944 considerava-se, de forma quase generalizada, que

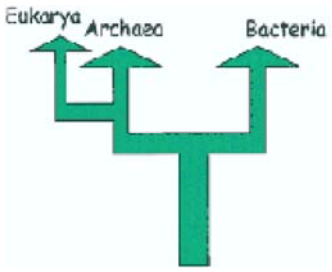


Figura 1.1: Domínios da vida pelos quais se dividem os organismos.

eram as proteínas cromossómicas as macromoléculas portadoras da informação genética. Comparado com a complexidade e diversidade das proteínas, o DNA parecia uma molécula simples demais para permitir explicar o conjunto de características específicas de todos os organismos. Os resultados das experiências realizadas com microrganismos vieram esclarecer a importância do DNA como material genético.
O DNA é uma biomolécula que pertence ao grupo dos ácidos nucleicos. Existem dois tipos de ácidos nucleicos:

- o ácido desoxirribonucleico (DNA);
- o ácido ribonucleico (RNA - Ribonucleic Acid);

e são polímeros em que as unidades básicas que os constituem são nucleótidos. Um nucleótido é constituído por três componentes essenciais (ver Figura 1.2):

- Ácido fosfórico - que lhes confere as suas características ácidas.
- Pentoses - que ocorrem na forma de dois tipos: a desoxirribose e a ribose. As designações destas pentoses relacionam-se com a existência de menos um átomo de oxigénio na desoxirribose do que na ribose.
- Bases azotadas - que formam dois grupos, dividindo as cinco bases azotadas existentes:
  - bases de anel duplo - Adenina (A) e Guanina (G);
  - bases de anel simples - Timina (T), Citosina (C) e Uracilo (U).

Os nucleótidos são designados pela base que entra na sua constituição. Assim, podem considerar-se cinco categorias de nucleótidos: nucleótido Adenina, nucleótido Guanina, nucleótido Citosina, nucleótido Timina e nucleótido Uracilo.
Em cada um dos ácidos nucleicos existem apenas quatro das cinco bases azotadas referidas:

- a Timina só existe no DNA;
- o Uracilo só existe no RNA;
- as outras três são comuns aos dois ácidos.

Os nucleótidos podem unir-se sequencialmente, constituindo uma cadeia polinucleotídica conforme mostra a Figura 1.3.
Observando a Figura 1.3 verifica-se que cada novo nucleótido liga-se pelo grupo fosfato ao

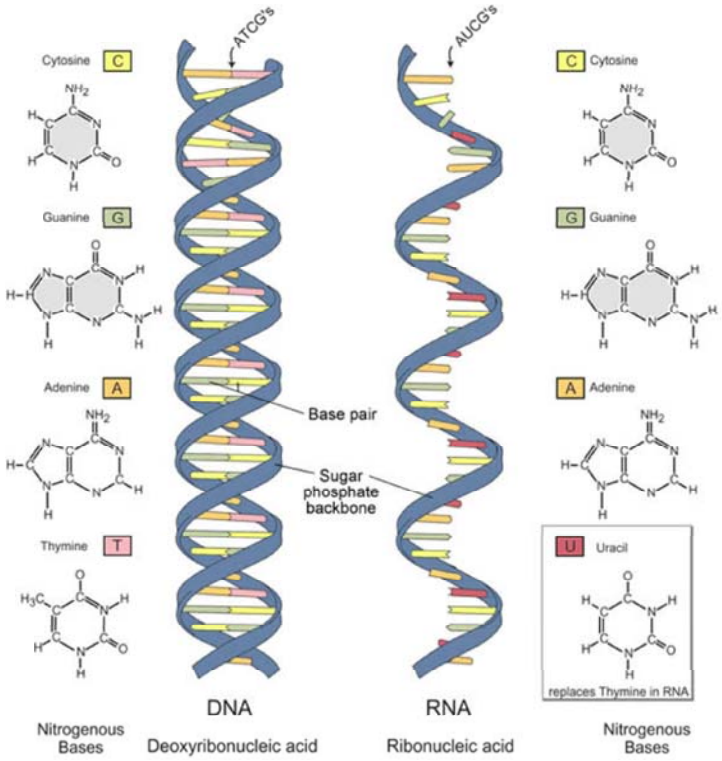


Figura 1.2: Estrutura do DNA e RNA.

carbono $3'$(C$3'$) da pentose do último nucleótido da cadeia, repetindo-se o processo no sentido do carbono $5'$ (C$5'$) para o C$3'$ ($5' \rightarrow 3'$). Assim, ao último nucleótido, que tem o C$3'$ livre, pode ligar-se um novo nucleótido pelo grupo fosfato.

J. Watson e F. Crick apresentaram, em 1953, a proposta de modelo dupla hélice para a estrutura do DNA, como é visível na Figura 1.2. Este modelo assemelha-se a uma escada de corda enrolada helicoidalmente em que:

- as bandas laterais da hélice são formadas por moléculas de fosfato, alternando com moléculas de desoxirribose, e os "degraus" centrais são pares de bases ligados entre si por ligações de hidrogénio;
- as bases que emparelham são bases complementares:
  - a Adenina liga-se, por ligação covalente dupla, à Timina (A=T);
  - a Guanina liga-se, por ligação covalente dupla, à Citosina (G=C);
- as duas cadeias polinucleotídicas de dupla hélice desenvolvem-se em direcções opostas. Cada uma delas inicia-se por uma extremidade de C$5'$ e termina em C$3'$. À extremidade

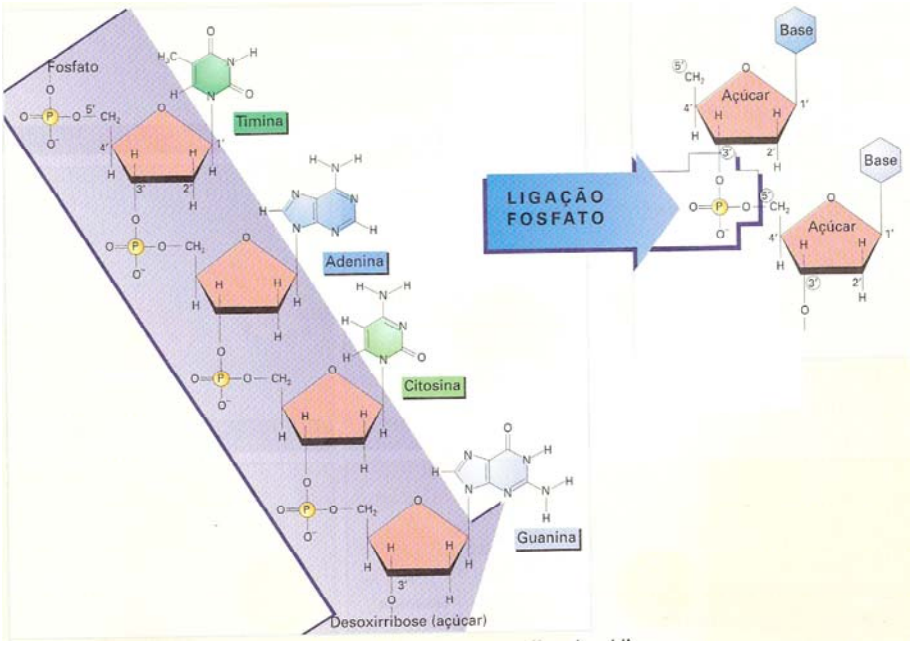


Figura 1.3: Ligação de um novo nucleótido.

C3′ de uma cadeia corresponde a extremidade C5′ da outra, designando-se por isso cadeias antiparalelas.

Um segmento da molécula de DNA denomina-se por gene. Cada gene pode conter milhares de nucleótidos. A grande diversidade de genes deve-se ao número de bases que cada um contém, à natureza de cada base e à ordem pela qual essas bases estão organizadas. Estes segmentos de informação genética controlam a síntese proteica.
Ao conjunto de genes, que constitui a informação genética de um individuo, dá-se o nome de genoma.

O outro ácido nucleico, o RNA, tem dimensões muito inferiores às do DNA. Ao nível da constituição, a grande diferença entre ambos, como já se referiu, é que no RNA a pentose é a ribose e nas bases azotadas troca a Timina pelo Uracilo (ver Figura 1.2). Conforme as funções que desempenha, o RNA pode ocorrer em formas estruturais diferentes. Em determinadas regiões a cadeia simples pode dobrar-se, e devido à complementaridade dos pares de bases, estabelecem-se ligações $A = U$ e $C = G$.

Os cientistas estabeleceram um código de correspondência entre os ácidos nucleicos e as proteínas, chamado *código genético*. O código genético funciona como um "dicionário" com-

posto por um alfabeto de 4 "letras" que a célula utiliza quando se dá a transferência da informação genética para a síntese das proteínas.

Nesse código cada três nucleótidos consecutivos de DNA codifica um aminoácido, unidade básica da proteína. Existem 64 (4^3) possibilidades de combinar três das quatro "letras", o que é suficiente para codificar os cerca de 20 aminoácidos (ver Figura 1.4). Cada tripleto de nucleótidos que codifica um aminoácido ou o início ou o fim da síntese proteica tem o nome de codão.

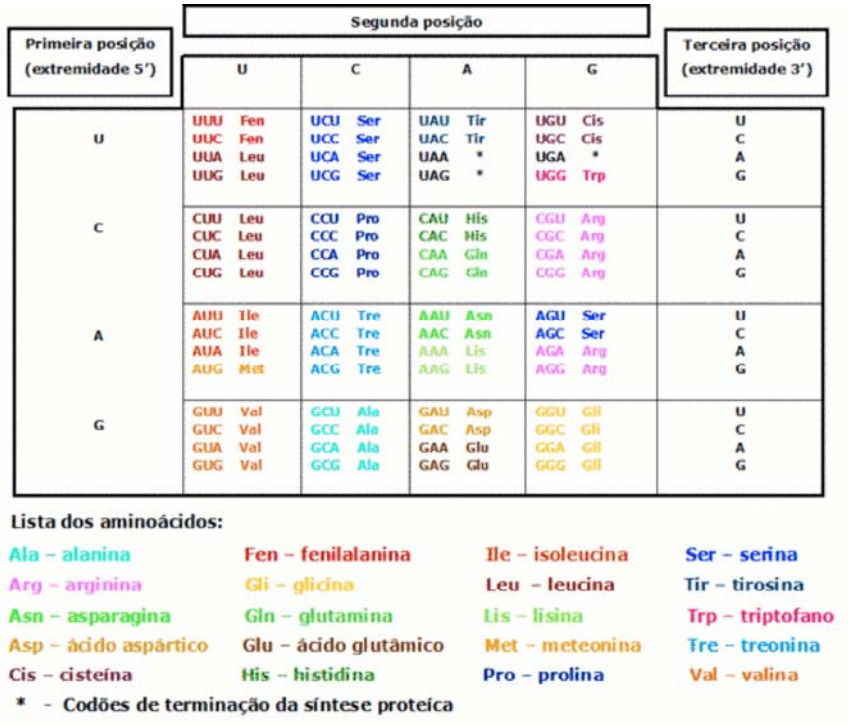


Figura 1.4: Tabela de correspondência entre codões e aminoácidos.

Vários dados relativos ao código genético permitiram identificar algumas das suas características:

- a "universalidade" do código genético - há uma linguagem comum a quase todas as células;
- a redundância - neste código vários codões são sinónimos, isto é, podem codificar o mesmo aminoácido;
- o código não é ambíguo - o mesmo codão, em regra, não codifica aminoácidos diferentes;
- o terceiro nucleótido de cada codão não é tão específico como os dois primeiros - por

exemplo, o aminoácido Arginina pode ser codificado pelos codões: CGU, CGC, CGA e CGG, confronte-se com a Figura 1.4;

- o tripleto AUG tem uma dupla função - codifica o aminoácido Meteonina e é o codão de iniciação da síntese proteica;
- os tripletos UUA, UAG, UGA são codões de finalização - estes codões, embora não codifiquem aminoácidos, representam o final da síntese proteica.

A informação contida no DNA é constituída por sequências que não codificam, chamadas intrões, intercaladas com sequências que codificam, designadas por exões. A transcrição de um segmento de DNA forma um RNA pré-mensageiro. No processamento deste RNA, por acção de enzimas, são retirados os intrões, havendo posteriormente a união dos exões. Estas transformações conduzem à formação do RNA mensageiro (mRNA), o qual abandona o núcleo, transportando a mensagem, ainda em código, para os ribossomas, onde a mensagem é descodificada, ou seja, traduzida para linguagem proteica (ver Figura 1.5). A

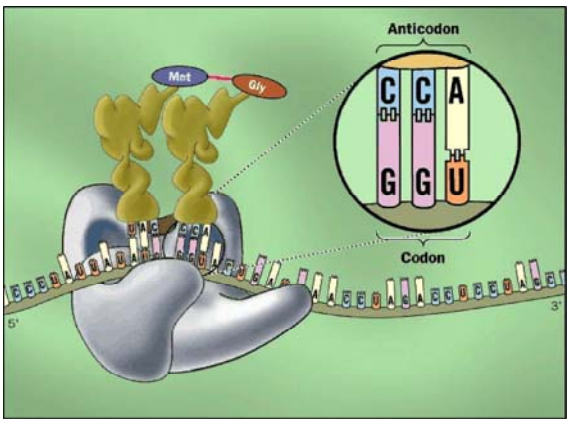


Figura 1.5: Construção da proteína a partir da leitura da sequência de codões pelo ribossoma.

tradução corresponde à transformação da mensagem contida no mRNA, através do RNA de transferência (tRNA). Em determinada região de cada tRNA existe uma sequência de 3 nucleótidos, chamada anticodão, que é complementar com um dos codões do mRNA. A cada tRNA, com determinado anticodão, irá corresponder uma ligação ao aminoácido respectivo, através de enzimas específicas.

O processo de leitura inicia-se com a ligação entre o mRNA e o tRNA iniciador, que transporta a Meteonina (ver Figura 1.6). Um novo tRNA, que transporta um segundo aminoácido, liga-se ao segundo codão, existindo a formação de uma primeira ligação peptídica entre o

aminoácido que ele transporta e a Meteonina. O ribossoma avança três bases, o processo repete-se ao longo do mRNA. Iniciada a leitura da molécula de mRNA, e dado um codão fixo da molécula, diz-se que o codão que o antecede está na posição $5'$ e o codão que o sucede está na posição $3'$. A leitura da molécula de mRNA pelo ribossoma é feita da posição $5'$ para a posiçao $3'$ (ver Figura 1.7). Na última fase, os codões de finalização (UAA, UAG, UGA), uma vez que não têm nenhum anticodão complementar, quando decifrados pelo ribossoma, indicam o fim da síntese proteica (ver Figura 1.8).

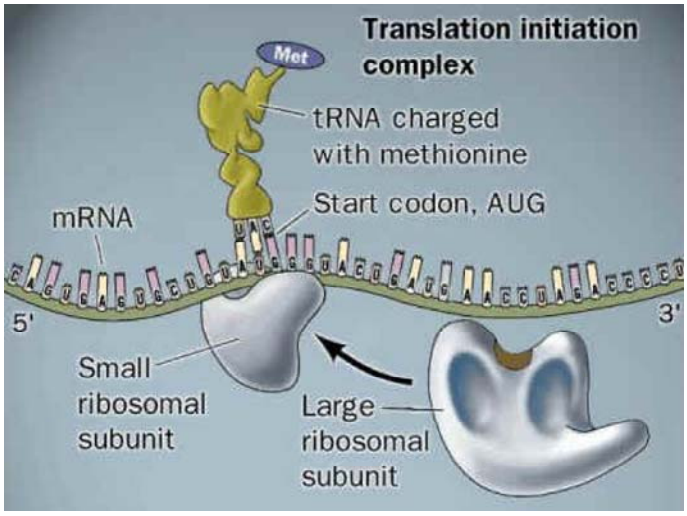


Figura 1.6: Processo de tradução da mensagem contida no mRNA.

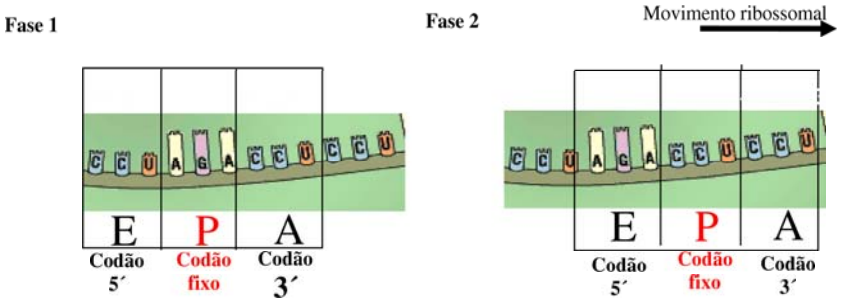


Figura 1.7: Esquema que explicita o sentido de leitura do ribossoma.

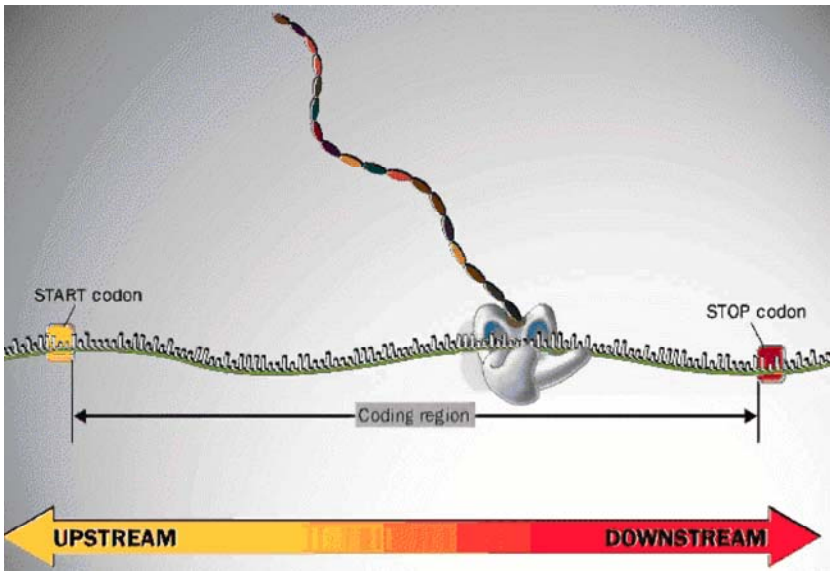


Figura 1.8: Esquema do processo de leitura da molécula de mRNA.

1.3 Preliminares

Perante a enorme quantidade de dados discretos resultantes da descodificação do código genético de um número cada vez maior de espécies, torna-se importante definir objectivos restritos, a fim de processar dados de modo conveniente à concretização desses mesmos objectivos. O propósito aqui será inferir sobre os parâmetros de amostras resultantes de dados genómicos. Ter-se-à em consideração a estimação pontual e intervalar dos parâmetros dos modelos teóricos assumidos. Por estimação pontual privilegia-se a precisão ao passo que por estimação intervalar, se dá maior relevância à confiança que se possa atribuir às estimativas propostas.

Uma estimativa pontual de um parâmetro desconhecido é um valor obtido a partir da amostra (através de uma estatística, função da amostra aleatória) que se destina a fornecer valores aproximados do parâmetro. Os estimadores considerados, para este trabalho, eram escolhidos de modo a satisfazer as propriedades de um "bom" estimador:

- Não enviesado - Um estimador, $\widehat{\theta}$, do parâmetro θ, é centrado ou não enviesado se

$$E(\widehat{\theta}) = \theta.$$

  Caso o estimador seja enviesado, a diferença *viés*$= E[\widehat{\theta}] - \theta$, mede o enviesamento do estimador;

- Eficiente - Entre os estimadores centrados, um estimador $\widehat{\theta_1}$ é mais eficiente que outro $\widehat{\theta_2}$ se

$$V[\widehat{\theta_1}] < V[\widehat{\theta_2}];$$

- Menor Erro Quadrático Médio (EQM) - O EQM de um estimador $\widehat{\theta}$ é dado por

$$EQM(\widehat{\theta}) = E[(\widehat{\theta} - \theta)^2] = V[\widehat{\theta}] + (E(\widehat{\theta} - \theta)^2).$$

Uma estimativa pontual de um parâmetro não contém informação sobre a precisão do valor obtido. Mas a variância e o EQM fornecem alguma informação sobre a variabilidade do estimador para o parâmetro. Assim, a forma mais habitual de inferir o valor de um parâmetro desconhecido é construir estimativas na forma de intervalo, identificando a probabilidade deste conter o verdadeiro valor do parâmetro. Um intervalo de confiança para um parâmetro θ, a um nível de confiança de $1-\alpha$, é uma estimativa de um intervalo aleatório (Θ_1, Θ_2) onde $P[\Theta_1 < \theta < \Theta_2] = 1-\alpha,\ \alpha \in (0,1)$.
Normalmente α é um valor estipulado à partida e é muito reduzido de forma a definir confiança elevada no intervalo estimado.
À diferença $\Theta_2 - \Theta_1$ chama-se amplitude ou largura do intervalo.
O ideal seria ter-se intervalos de amplitude reduzida e confiança elevada pois assim ficaríamos a conhecer os parâmetros da distribuição com boa precisão.

1.4 Motivação e Organização da Dissertação

A presente dissertação de Mestrado surge integrada num projecto interdisciplinar de investigação com o objectivo geral e global de contribuir para a compreensão da estrutura da linguagem genética ou códigos genéticos.
Para além deste objectivo, pretende-se também contemplar a componente Ensino apresentando instrumentos de trabalho relacionados com as distribuições discretas e associados a conteúdos programáticos leccionados na disciplina de Matemática A no Ensino Secundário.
Serão alvo de estudo as sequências codificantes de código genético de várias espécies dos domínios da Vida: Archaea, Eukarya e Bacteria, considerando como principal objectivo a estimação de parâmetros de distribuições associadas ao comportamento de pares de codões iguais e a codões cujas frequências na sequência de DNA é muito baixa, sendo estes designados por codões raros.
Segundo os biólogos, o genoma de uma dada espécie é representativo de todos os indivíduos

pertencentes a essa espécie. Partindo deste princípio, as conclusões que se tirarem para o genoma observado de uma espécie são extensivas a todos os indivíduos dessa espécie. Para todos os genes de um genoma são retiradas informações relativas aos codões sequenciados pela ordem de leitura feita no ribossoma. Esta enorme quantidade de informação, assim obtida das sequências de código genético, foi organizada em tabelas de contingência de pares de codões justapostos segundo as leituras $3'$ e $5'$; nesta investigação considerou-se somente a leitura $3'$ por ser a mais natural.

Pelo facto de entre uma cadeia simples de DNA e uma cadeia de mRNA existir uma aplicação bijectiva entre nucleótidos, as leis que se obtêm para o sequenciamento de codões no DNA têm correspondência imediata para a sequência de codões no mRNA. De agora em diante far-se-à apenas referência ao mRNA dado que se pretende apenas decifrar leis que regulam a tradução do mRNA pelo ribossoma.

Esta dissertação é constituída, para além deste capítulo introdutório, por mais cinco capítulos e diversos apêndices.

No capítulo seguinte serão abordados os desenvolvimentos teóricos de metodologias estatísticas apropriadas à análise de dados discretos, resultantes de populações com distribuições Binomial, Multinomial e Poisson.

No Capítulo 3 analisam-se e comparam-se vários procedimentos que garantem a significância global fixada na realização de testes simultâneos em tabelas de contingência, permitindo assim estabelecer diferentes formas de identificar pares de codões problemáticos no sequenciamento do mRNA.

No Capítulo 4 apresenta-se uma aplicação das metodologias abordadas na análise estatística de dados genómicos para identificar regras que regulam a associação de pares de codões iguais e avaliar modelos probabilísticos de codões raros, em organismos dos três domínios da Vida.

No Capítulo 5, e no âmbito do Tema: Probabilidades e Análise Combinatória do programa oficial de Matemática A, 12.º ano, do Ensino Secundário, apresenta-se material didáctico e pedagógico com o objectivo de introduzir propriedades associadas às distribuições discretas de Bernoulli, Binomial, Hipergeométrica e Poisson, bem como a distribuição contínua Normal. Para tal construíram-se aplicações interactivas com o intuito de:

- analisar a variação dos parâmetros e respectiva influência no comportamento das funções de massa de probabilidade e função de densidade de probabilidade;

- estabelecer, de modo intuitivo, regras de convergência da Binomial à Normal, da Hipergeométrica à Binomial e desta à distribuição de Poisson.

O Capítulo 6 conclui esta dissertação, resumindo os resultados obtidos nas aplicações realizadas no Capítulo 4, deixando-se em aberto algumas ideias e direcções de trabalho para futuras investigações.

Capítulo 2

Distribuições Discretas

2.1 Distribuição Binomial

A distribuição Binomial é uma das mais antigas que se conhecem e é o modelo probabilístico adequado para descrever fenómenos associados a sucessões de experiências aleatórias independentes em que, em cada uma, se observa a ocorrência ou não de determinado acontecimento, de probabilidade constante de experiência para experiência.

2.1.1 Definição e Propriedades

A distribuição de Bernoulli é utilizada para modelar variáveis que só podem assumir dois valores distintos.

A notação $Y \sim B(p)$ significa que a variável aleatória (v.a.) discreta Y segue uma distribuição de Bernoulli, com parâmetro p, atribuindo-se os valores 0 ou 1 à variável e com função de massa de probabilidade (fmp)

$$P[Y = j] = p^j (1-p)^{1-j}, \quad j = 0, 1, \ p \in (0, 1).$$

O valor médio e a variância são, respectivamente, $E[Y] = p$ e $V[Y] = p(1-p)$.

Na Figura 2.1 ilustra-se a fmp de uma distribuição de Bernoulli.

A distribuição Binomial é uma generalização da distribuição de Bernoulli quando se considera o número de sucessos em n provas independentes de Bernoulli.

A notação $Y \sim B(n, p)$ significa que a v.a. discreta Y segue uma distribuição Binomial baseada em n provas independentes com igual probabilidade de sucesso p. A sua fmp é

$$P[Y = j] = \binom{n}{j} p^j (1-p)^{n-j}, \quad j = 0(1)n, \ p \in (0, 1), \tag{2.1}$$

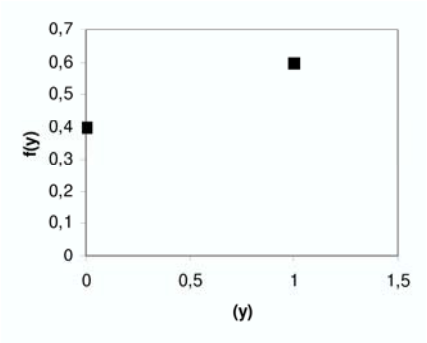


Figura 2.1: A fmp de uma variável com distribuição B($p = 0.6$).

onde $\binom{n}{j}$ representa as combinações de n elementos, j a j.
O valor médio e a variância são $E[Y] = np$ e $V[Y] = np(1-p)$, respectivamente.
Na Figura 2.2 visualiza-se como variam os gráficos da fmp da distribuição Binomial em relação ao parâmetro n: o aumento do parâmetro faz com que a forma da distribuição se aproxime mais a um sino.

Quanto ao parâmetro p, a Figura 2.3 ilustra a sua influência na forma da distribuição: valores

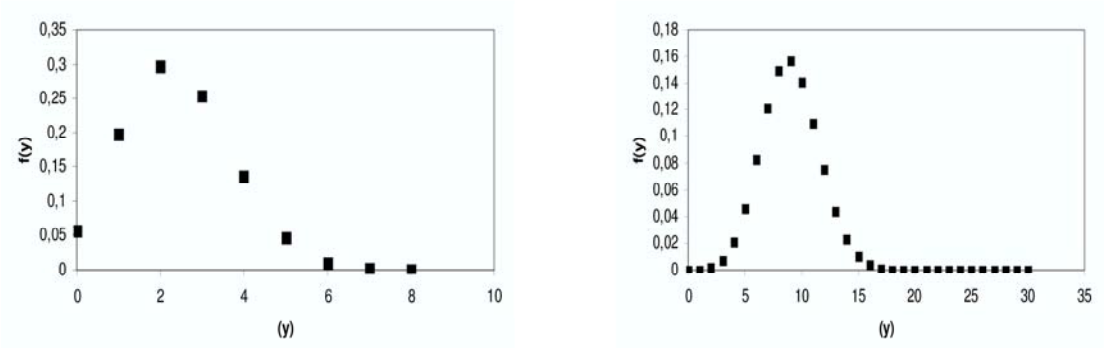


Figura 2.2: As fmp′s de variáveis com distribuição B(8,0.3) e B(30,0.3).

inferiores a 0.5 produzem assimetria positiva, valores superiores a 0.5 produzem assimetria negativa, e quando $p = 0.5$ a distribuição é simétrica.

Se $Y_1 \sim B(n_1, p)$ é independente de $Y_2 \sim B(n_2, p)$ então, para $0 \leq t \leq n_1 + n_2$ fixo, tem-se

$$P[Y_1 = j | Y_1 + Y_2 = t] = \frac{\binom{n_1}{j}\binom{n_2}{t-j}}{\binom{n_1+n_2}{t}}, \quad \max\{0, t-n_2\} \leq j \leq \min\{n_1, t\},$$

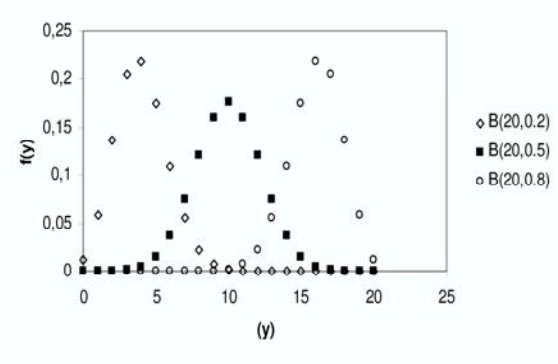


Figura 2.3: As fmp's de variáveis com distribuições B(20,0.2), B(20,0.5) e B(20,0.8).

que representa a fmp da distribuição de probabilidade conhecida por distribuição Hipergeométrica, denotada por $H(N, M, n)$, de parâmetros $N = n_1 + n_2$, $M = n_1$ e $n = t$.
Assumindo $Y \sim B(n, p)$, o Teorema de Limite Central (TLC) permite estabelecer aproximações da função de distribuição (f.d.) de Y, sob normalização linear conveniente, para valores "elevados" de n. Tem-se

$$P[Y \leq j] \approx \Phi\left(\frac{j - np}{\sqrt{np(1-p)}}\right), \qquad n \to \infty, \tag{2.2}$$

ou com o factor de correcção à continuidade

$$P[Y \leq j] \approx \Phi\left(\frac{j + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right), \qquad n \to \infty, \tag{2.3}$$

onde Φ é a função de distribuição da distribuição Normal *standard*, $N(0, 1)$.
A aproximação (2.2) é indicada para valores centrais de p e valores "elevados" de n, ie, $0.1 \leq p \leq 0.9$ e $n \geq 20$ ([41]). Para ilustrar, observe-se que quando $Y \sim B(30, 0.1)$, vem

$$\begin{aligned} P[Y \leq 4] &= 0.825 \qquad \text{(valor exacto)} \\ &\approx 0.7286 \qquad \text{(a partir de 2.2)} \end{aligned}$$

$$\begin{aligned} P[Y \leq 4] &= 0.825 \qquad \text{(valor exacto)} \\ &\approx 0.8193 \qquad \text{(a partir de 2.3)} \end{aligned}$$

Pela aproximação (2.2) tem-se 11.6% de erro relativo contra 0.6% de erro relativo dado pela aproximação (2.3).

2.1.2 Estimação Pontual de p

Considere-se $Y \sim B(n,p)$ onde n é conhecido e p é desconhecido com $0 < p < 1$.
A estimativa pontual $\widehat{p} = \frac{Y}{n}$ é o estimador de máxima verosimilhança de p.
Uma aplicação da desigualdade de Cramér-Rao mostra que $\widehat{p}$ é o estimador Não Enviesado com Variância Uniformemente Mínima (UMVU) de p ([50]), ou seja

- $E(\widehat{p}) = p$;
- $V(\widehat{p}) = \frac{p(1-p)}{n} = EQM(\widehat{p})$.

2.1.3 Estimação Intervalar para p

Suponha-se que $Y \sim B(n,p)$ com p desconhecido.
Vários autores consideram diferentes procedimentos para construir intervalos de confiança bilaterais para o parâmetro p quer com base em amostras de pequenas quer de grandes dimensões. Intervalos unilaterais para pequenas amostras podem ser obtidos a partir de Clopper e Pearson ([13]); intervalos unilaterais para amostras de grandes dimensões são discutidos em Fujino ([22]) e Blyth ([7]).

Intervalos de Confiança para amostras de pequenas dimensões

Existe uma relação dual entre testes estatísticos e intervalos de confiança (IC). Pode sempre obter-se um IC para um parâmetro de interesse invertendo a família de testes.
Um intervalo de confiança para p, a um nível de confiança $(1-\alpha)100\%$, pode ser construído tendo em conta a região de aceitação associada ao teste de hipóteses bilateral, de dimensão menor ou igual a α, para testar

$$H_0 : p = p_0 \qquad vs \qquad H_1 : p \neq p_0 \qquad (\text{com } \ p_0 \in (0,1)). \tag{2.4}$$

Formalmente, definindo $A(p_0)$ como a região de aceitação de H_0, a um nível de significância inferior a α então

$$P_{p_0}[Y \in A(p_0)] = P[Y \in A(p_0) \mid H_0 : p = p_0] \geq 1 - \alpha,$$

e portanto, com base na região $A(p_0)$ são definidos IC para o parâmetro p. Observe-se que

$$P_{p_0}[p_0 \in \{p_0 : Y \in A(p_0)\}] = P_{p_0}[Y \in A(p_0)] \geq 1 - \alpha,$$

donde $\{p_0 : Y \in A(p_0)\}$ é o conjunto dos valores de p para um nível de confiança $(1-\alpha)100\%$.

Intervalos Centrados Uniformemente Exactos (ICUE)

Blyth e Hutchinson ([8]) construíram tabelas para intervalos correspondentes às regiões de aceitação $A(p_0)$ de testes centrados uniformemente mais potentes para testar (2.4). Estas regiões de aceitação $A(p_0)$ são definidas por um conjunto de números $\{L(p_0), ..., U(p_0)\}$, com extremos aleatórios, tal que cada "cauda" da região tem probabilidade $\frac{1}{\alpha}$. No entanto, não são usados na prática dada a sua aleatoriedade.

Intervalos de Caudas

Clopper e Pearson ([13]) propuseram um conjunto de intervalos não aleatórios com a tentativa de possuirem caudas-iguais.
Fixe-se $0 < \alpha < 1$ e defina-se a região de rejeição

$$R(p_0) := \underline{R}(p_0) \cup \overline{R}(p_0)$$

onde

$$\underline{R}(p_0) := \{\{0, ..., L\} : P_{p_0}[Y \leq L] \leq \tfrac{\alpha}{2} < P_{p_0}[Y \leq L + 1]\}$$

e

$$\overline{R}(p_0) := \{\{U, ..., n\} : P_{p_0}[Y \geq U] \leq \tfrac{\alpha}{2} < P_{p_0}[Y \leq U - 1]\}$$

Por palavras, as caudas inferior e superior, $\overline{R}(p_0)$ e $\underline{R}(p_0)$, do teste (2.4) de $H_0 : p = p_0$ tem probabilidade tão próxima de $\frac{\alpha}{2}$ quanto possível mas menor ou igual do que $\frac{\alpha}{2}$.
Note-se que, $\underline{R}(p_0)$ ($\overline{R}(p_0)$) pode ser vazio pois $Pp_0[Y = 0]$ ($Pp_0[Y = n]$) pode ser maior do que $\frac{\alpha}{2}$. Tal facto, conduz sempre à decisões de aceitação de H_0, levando a testes conservativos e, portanto, longos IC sem interesse.
Segundo, Clopper e Pearson, definem-se as regiões de aceitação como sendo

$$A(p_0) := \{0, ..., n\} \setminus R(p_0).$$

Assim, $A(p_0)$ é da forma $\{L(p_0), ..., U(p_0)\}$ e tem probabilidade de cobertura no mínimo $1-\alpha$, já que nestas condições,

$$P_{p_0}[Y \in A(p_0)] = 1-P_{p_0}[Y \in R(p_0)] \geq 1-P_{p_0}[Y \in \underline{R}(p_0)]-P_{p_0}[Y \in \overline{R}(p_0)] \geq 1-\frac{\alpha}{2}-\frac{\alpha}{2} = 1-\alpha.$$

O IC de Clopper e Pearson, a $(1 - \alpha)100\%$ para p, resultante da inversão de

$$\{A(p_0) : 0 < p_0 < n\}$$

e designados por intervalos de caudas, é dado por $I(j) := \{p_0 : j \in A(p_0)\} = (\underline{p}(j), \overline{p}(j))$, onde os valores $\underline{p}(j)$ e $\overline{p}(j)$ são calculados de modo que $\underline{p}(0) = 0, P_{\underline{p}(j)}[Y \geq j] = \frac{\alpha}{2}$, para $1 \leq j \leq n$, e $\overline{p}(n) = 1, P_{\overline{p}(j)}[Y \leq j] = \frac{\alpha}{2}$, para $0 \leq j \leq n$. O cálculo dos valores extremos do intervalo pode ser simplificado através das fórmulas,

$$\begin{aligned} \underline{p}(j) &= \frac{1}{1+\frac{(n-j+1)}{j} F_{\frac{\alpha}{2},2(n-j+1),2j}}, 1 \leq j \leq n \\ \overline{p}(j) &= \frac{\frac{j+1}{n-j} F_{\frac{\alpha}{2},2(j+1),2(n-j)}}{1+\frac{(j+1)}{n-j} F_{\frac{\alpha}{2},2(j+1),2(n-j)}}, 0 \leq j \leq n-1 \end{aligned} \tag{2.5}$$

onde F_{α,ν_1,ν_2} denota o quantil de ordem $1-\alpha$ da distribuição de Fisher com ν_1 e ν_2 graus de liberdade como indicam Santner e Duffy ([50]). Os intervalos de caudas têm muitas propriedades interessantes:

i) Se $(\underline{p}, \overline{p})$ é um IC para Y então $(1-\overline{p}, 1-\underline{p})$ é um IC para $n-Y$;

ii) É simetrico à volta de $1/2$ quando n é par e $j = n/2$, ou seja, $\underline{p}(n/2) = 1 - \overline{p}(n/2)$;

iii) Para n fixo, $\underline{p}(j) \leq \underline{p}(j+1)$ e $\overline{p}(j) \leq \overline{p}(j+1)$, para $0 \leq j \leq n-1$;

iv) Para $0 \leq j \leq n$ fixo, $\overline{p}(j)$ é não crescente em n;

v) Para $0 \leq j \leq n$, com n fixo, $\underline{p}(j)$ é não decrescente em α e $\overline{p}(j)$ é não crescente em α.

Os intervalos de caudas de Clopper e Pearson têm sido adaptados em muitos problemas dado o seu aspecto intuitivo e facilidade computacional. Porém são extremamente conservativos porque a região de rejeição $R(p_0) = \underline{R}(p_0) \cup \overline{R}(p_0)$ é muitas vezes demasiado pequena.
Outros autores (Vos ([61],[62]), Fujino e Okuno ([23])) proposeram outros sistemas de IC para p que podem ter interpretação Bayesiana.
Outros métodos para pequenas amostras trabalham directamente com regiões de aceitação para forçar a estar mais próximo do nível nominal $1-\alpha$. A descrição que se segue é uma dessas construções.

Intervalos de Confiança de Sterne/Crow/Blyth/Still

Diversas sugestões têm sido feitas na literatura para construir intervalos de caudas menos conservativos.
Sterne ([54]) estabelece um novo sistema de intervalos de caudas para pequenas amostras, tendo como base a ideia de que pequenos intervalos devem resultar de pequenas regiões de

aceitação. Aquele autor propôs que se pusesse os "mais prováveis" resultados em $A(p_0)$.
Formalmente, fixa-se um $\alpha \in (0,1)$ e $p_0 \in (0,1)$.
Denote-se por $l_1, ..., l_{n+1}$, os $n+1$ resultados possíveis $0, 1, ..., n$, da v.a. $Y \sim B(n,p)$ ordenados segundo o critério

$$P_{p_0}[Y = l_1] \geq P_{p_0}[Y = l_2] \geq \ ... \geq P_{p_0}[Y = l_{n+1}].$$

Define-se a região de aceitação da hipótese nula $H_0 : p = p_0$ por $A(p_0) := \{l_1, l_2, ..., l_k\}$ onde k satisfaz

$$\sum_{i=1}^{k} P_{p_0}[Y = l_i] \geq 1 - \alpha > \sum_{i=1}^{k-1} P_{p_0}[Y = l_i], \tag{2.6}$$

por forma a que $A(p_0)$ contenha os resultados mais prováveis e em menor número, de modo a conseguir a desejada probabilidade de cobertura $1-\alpha$. Assim, $A(p_0)$ é de menor cardinalidade possível e, portanto, uma pequena região $A(p_0)$ gerará conjuntos de confiança mais pequenos.
Dois problemas estão associados às regiões de aceitação consideradas por Sterne:

- o primeiro é a ambiguidade na definição de $A(p_0)$. Fixando n, α e p_0, podem existir dois valores distintos, l_j e l_{j+1}, tais que $P_{p_0}[Y = l_j] = P_{p_0}[Y = l_{j+1}]$ conduzindo os dois resultados à definição de uma região de aceitação $A(p_0)$ satisfazendo (2.6) mas não de modo único;

- o segundo problema é o facto de como Crow ([18]) apontou os conjuntos de confiança de Sterne, $I(j) = \{p_0 : j \in A(p_0)\}$, podem não ser necessariamente intervalos.

A Tabela 2.1 lista regiões de aceitação $A(p_0) = \{l_1, l_2, \cdots, l_k\}$ a 95% de confiança, para $n = 25$ e vários valores de p_0. Invertendo os conjuntos de confiança, obtém-se, por exemplo, para $y = 0$ o intervalo de Sterne, $I(0) = (0, 0.133)$.
Crow ([18]) propôs um algoritmo de construção dos conjuntos de aceitação

$$A(p_i) = \{L(p_i), ..., U(p_i)\}$$

para a partição $0 < p_1 < ... < p_M < 1$ de $[0,1]$ para forçar que $L(p_i)$ e $U(p_i)$ sejam não decrescentes. O que não acontece com os conjuntos de Sterne.
Simultaneamente, Crow atendeu a forçar $A(p_i)$ a conter o mesmo número de pontos do que as regiões de aceitação de Sterne. Consequentemente, os intervalos de Crow minimizam (2.6) na maior parte dos casos.
Blyth e Still ([9]) notaram que os intervalos de Crow não respeitam a maioria das propriedades referidas atrás, e por isso, construíram um conjunto de intervalos de Crow modificados que já

p_0	$A(p_0)$
0.001	$\{0\}$
0.006	$\{0,1\}$
0.127	$\{0,...,6\}$
0.128	$\{0,...,6\}$
0.133	$\{0,...,6\}$
0.134	$\{1,...,7\}$
0.135	$\{1,...,7\}$

Tabela 2.1: Regiões de aceitação de Sterne $A(p_0)$ a 95% de confiança, para $n = 20$ e vários valores de p_0.

satisfazem essas propriedades, contendo o mesmo número de pontos que a região de aceitação $A(p_0)$ de Sterne para (j, n, α).

Casella ([12]) determina uma classe completa de intervalos de confiança invariantes para p a um nível de confiança $(1-\alpha)100\%$ para $\alpha = 0.01, 0.05$ e $n = 1(1)30$. Os intervalos de Blyth e Still são membros desta classe. Os extremos destes intervalos podem ser obtidos recorrendo a tabelas [9].

Intervalos de Confiança para amostras de grandes dimensões

Sendo $Y \sim B(n,p)$, o TLC permite estabelecer

$$\frac{\sqrt{n}(\widehat{p}-p)}{\sqrt{p(1-p)}} \sim N(0,1).$$

Substituindo p por $\widehat{p}$ no denominador obtém-se uma das formas mais usadas para obter uma estimativa intervalar para p, a um nível de confiança $(1-\alpha)100\%$ e quando a amostra é de grandes dimensões, dada por

$$I_N = \left(\widehat{p} - c\sqrt{\widehat{p}(1-\widehat{p})},\ \widehat{p} + c\sqrt{\widehat{p}(1-\widehat{p})}\right),$$

onde $c := z_{\alpha/2}/\sqrt{n}$ e $z_{\alpha/2}$ é o quantil de ordem $1-\alpha/2$ da distribuição Normal *standard*.

Uma versão mais sofisticada daquela estimativa intervalar para p é dada por

$$I_S = \left(\frac{2\widehat{p} + c^2 - c\sqrt{c^2 + 4\widehat{p}(1-\widehat{p})}}{2(1+c^2)}, \frac{2\widehat{p} + c^2 + c\sqrt{c^2 + 4\widehat{p}(1-\widehat{p})}}{2(1+c^2)}\right)$$

onde os limites de I_S são as soluções da equação em p: $\left(\frac{\sqrt{n}(\widehat{p}-p)}{\sqrt{p(1-p)}}\right)^2 = z^2_{\alpha/2}$.

Observe-se que I_N estima o erro *standard* de $\widehat{p}$ ao passo que em I_S se encontra o seu valor

exacto.

Ghost([25]) comparou I_S e I_N tendo observado que

- a probabilidade de cobertura de I_S é maior que a probabilidade de cobertura de I_N estando a primeira próxima do valor nominal $1-\alpha$;
- as probabilidades de cobertura de I_N são inadequadas para p próximo de 0 ou 1.
- I_S tende a ter uma amplitude menor e menos enviesado do que I_N.

Assim, do trabalho de Ghost, conclui-se que I_S apresenta melhores características em termos de probabilidade de cobertura, amplitude e viés do que I_N.

Diversas modificações de I_S, que enriqueceram ainda mais as suas características, foram propostas na literatura. Blyth e Still ([9]) propuseram um intervalo com continuidade corrigida, I_{BS}, definido por

$$I_{BS} = \left(\frac{n\widehat{p} - 0.5 + z^2_{\alpha/2}/2 - z_{\alpha/2}\sqrt{n\widehat{p} - 0.5 - (n\widehat{p} - 0.5)^2/n + z^2_{\alpha/2}/4}}{n + z^2_{\alpha/2}} \right. \\ \left. \frac{n\widehat{p} + 0.5 + z^2_{\alpha/2}/2 + z_{\alpha/2}\sqrt{n\widehat{p} + 0.5 - (n\widehat{p} - 0.5)^2/n + z^2_{\alpha/2}/4}}{n + z^2_{\alpha/2}} \right) \tag{2.7}$$

tendo-se observado que o intervalo I_{BS} consegue alcançar um nível de confiança próximo de $(1-\alpha)100\%$ para valores de n não inferiores a 30.

2.1.4 Estimação intervalar para w

Em vez de estimar a probabilidade de sucesso p, é muitas vezes conveniente pensar em termos da razão entre o sucesso e a falha, definindo o parâmetro chance de sucesso (em inglês *odds of sucess*):

$$w = p/(1-p).$$

O valor de w varia de 0 a ∞. Quando $p = 0$ vem que $w = 0$, quando $p \to 1$ significa que a ocorrência de sucesso tende a ser completamente garantida, e portanto, a chance de sucesso, w, vai tender para um valor infinitamente grande. Se w admite, por exemplo, um valor igual a 5 significa que $p = 5(1-p)$, logo w indicará que a probabilidade de ocorrência de sucesso é 5 vezes superior à probabilidade de ocorrência de insucesso.

A Figura 2.4 mostra que w como uma função de p, é estritamente crescente. Tal facto, leva

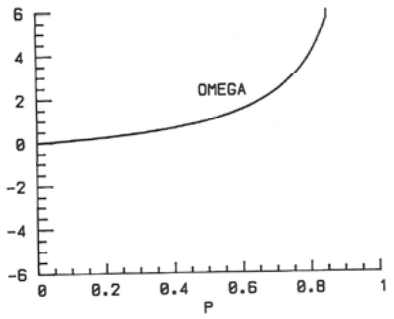


Figura 2.4: Representação gráfica de w *vs* p.

a que estimativas pontuais e intervalares para w possam ser determinadas directamente de estimativas de p. Concretamente, se $\widehat{p}$ é a estimativa de máxima verosimilhança de p então uma estimativa pontual para w é dada por $\frac{\widehat{p}}{1-\widehat{p}}$; se é conhecida uma estimativa intervalar para p, com extremos inferior e superior $\widehat{\theta}_1$ e $\widehat{\theta}_2$, respectivamente, uma estimativa intervalar para w será dada por

$$\left(\frac{\widehat{\theta}_1}{1-\widehat{\theta}_1} \, , \, \frac{\widehat{\theta}_2}{1-\widehat{\theta}_2} \right).$$

2.1.5 Estimação intervalar para $p_1 - p_2$

Suponhamos que X e Y são duas v.a. independentes extraídas de duas populações diferentes com $X \sim B(m, p_1)$ e $Y \sim B(n, p_2)$. O objectivo é comparar a diferença entre as probabilidades de sucesso destas duas populações, $p_1 - p_2$.

Várias formas de estabelecer IC, a um nível de confiança $(1-\alpha)100\%$, para $p_1 - p_2$, são conhecidas.

Denote-se por $\widehat{p}_1 = X/m$ e $\widehat{p}_2 = Y/n$ os estimadores de máxima verosimilhança de p_1 e p_2, respectivamente, e seja $z_{\alpha/2}$ o quantil de ordem $1-\alpha/2$ da distribuição Normal *standard*.

Intervalo de confiança *standard* (Wald)

Do TLC resulta a aproximação

$$\widehat{p}_1 - \widehat{p}_2 = \frac{X}{m} - \frac{Y}{n} \approx N(p_1 - p_2, V(\widehat{p}_1 - \widehat{p}_2)), \quad m, n \to \infty,$$

com $V(\widehat{p}_1) = \frac{p_1(1-p_1)}{m}$ e $V(\widehat{p}_2) = \frac{p_2(1-p_2)}{n}$.

Estimando $V(\widehat{p}_1 - \widehat{p}_2)$ por $\frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}$, obtém-se um IC para $p_1 - p_2$ a um nível de

confiança $(1-\alpha)100\%$ e dado por

$$\left(\widehat{p}_1 - \widehat{p}_2 \pm z_{\alpha/2}\sqrt{\widehat{p}_1(1-\widehat{p}_1)/m + \widehat{p}_2(1-\widehat{p}_2)/n}\right).$$

O intervalo de confiança de Wald é o mais intuitivo, e tendo uma forma simples, é muito utilizado na literatura.

Intervalo de confiança de Yule

Este IC é obtido a partir da suposição de que $p_1 = p_2 = p$. Assim,

$$V(\widehat{p}_1 - \widehat{p}_2) = \left(\frac{1}{m} + \frac{1}{n}\right) p(1-p),$$

sendo então $\overline{p} = (X+Y)/(m+n)$ um melhor estimador de $p = p_1 = p_2$.
O IC de Yule para $p_1 - p_2$, a um nível de confiança $(1-\alpha)100\%$, é dado por

$$\left(\widehat{p}_1 - \widehat{p}_2 \pm z_{\alpha/2}\sqrt{(1/m + 1/n)\overline{p}(1-\overline{p})}\right).$$

Apesar do intervalo de confiança de Yule ser derivado partindo do pressuposto que $p_1 = p_2$, mostra-se que tem um comportamento relativamente bom quando $|p_1 - p_2|$ não é muito grande, especialmente quando $m \approx n$ ([10]).

Intervalo de confiança modificado de Yule

O intervalo de confiança de Yule tem um bom comportamento quando $m = n$. Se $m \neq n$, nota-se um significante desvio da probabilidade de cobertura a partir do nível nominal quando p_1 e p_2 estão próximos de 0 ou 1. Então é necessário fazer alguma modificação.
Usando o estimador $\check{p} = (nX/m + mY/n)/(m+n)$ em vez de $\overline{p}$ obtem-se o intervalo de confiança modificado de Yule:

$$\left(\widehat{p}_1 - \widehat{p}_2 \pm z_{\alpha/2}\sqrt{(1/m + 1/n)\check{p}(1-\check{p})}\right)$$

Este procedimento produz enviesamento nos IC mais pequeno quando $m \neq n$, comparativamente com o IC de Yule. De notar, que quando $m = n$ o intervalo de confiança de Yule é um caso particular do modificado de Yule. Em ambos, o erro *standard* converge para o verdadeiro valor se e só se $p_1 = p_2$ à medida que $\max(m,n) \to \infty$.

Intervalo de Newcombe

Usando a informação a partir das amostras individuais dos intervalos *score* para p_1 e p_2,

Newcombe ([43]) propôs um intervalo híbrido.

Diz-se que (l_i, u_i) é o intervalo *score* para p_i se os extremos, $l_i < u_i$, são as raízes reais das equações quadráticas $z_{\alpha/2} = (\widehat{p}_i - p_i/\sqrt{p_i(1-p_i)/n_i})$, $i = 1, 2$, sendo $n_1 = m$, $n_2 = n$.

O intervalo de Newcombe para $p_1 - p_2$ resulta dos intervalos de scores e tem a forma

$$\left(\widehat{p}_1 - \widehat{p}_2 - z_{\alpha/2}\sqrt{\frac{l_1(1-l_1)}{m} + \frac{u_2(1-u_2)}{n}}, \widehat{p}_1 - \widehat{p}_2 + z_{\alpha/2}\sqrt{\frac{u_1(1-u_1)}{m} + \frac{l_2(1-l_2)}{n}}\right).$$

Intervalo de confiança recentrado

Para construir o intervalo de confiança de Wald para o parâmetro p, com base numa amostra de uma população com distribuição Binomial, simplesmente inverte-se o teste resolvendo a equação $(\widehat{p}-p)/\sqrt{p(1-p)/n} = z_{\alpha/2}$ em p. Para a diferença de proporções, esse procedimento não é aplicável, uma vez que existe um parâmetro perturbador p_1 (ou equivalentemente p_2). No entanto, com a reparametrização proposta por Brown e Li ([10]), e alguma aproximação razoável, é possível construir um teste de hipótese para testar $H_0 : p_1 - p_2 = 0$ e invertê-lo por forma a definir um intervalo de confiança para $p_1 - p_2$.

Sem perda de generalidade, assuma-se que $(p_1 - p_2) \geq 0$ e defina-se um novo parâmetro $\pi = (np_1 + mp_2)/(m+n)$. Assim, $p_1 = \pi + (p_1 - p_2)m/(m+n)$ e $p_2 = \pi - (p_1 - p_2)n/(m+n)$. Um estimador natural de π é

$$\widehat{\pi} = \frac{n\widehat{p}_1 + m\widehat{p}_2}{m+n}.$$

Observe-se que

$$V(\widehat{p}_1 - \widehat{p}_2) = \left(\frac{1}{m} + \frac{1}{n}\right)\pi(1-\pi) - \frac{(p_1 - p_2)^2}{m+n}.$$

Consequentemente, um estimador para $V(\widehat{p}_1 - \widehat{p}_2)$ é

$$\widehat{\sigma^2} = \left(\frac{1}{m} + \frac{1}{n}\right)\widehat{\pi}(1-\widehat{\pi}) - \frac{(\widehat{p}_1 - \widehat{p}_2)^2}{m+n}.$$

Para testar $H_0 : p_1 - p_2 = 0$, Brown e Li definem a região de rejeição através da condição $|(\widehat{p}_1 - \widehat{p}_2) - (p_1 - p_2)| \geq k\widehat{\sigma}$, onde k é o quantil de ordem $1 - \alpha/2$ da distribuição T de *Student* com $(m+n-2)$ graus de liberdade. Esses autores não apresentam justificações teóricas para a utilização do quantil da distribuição T de *Student* em vez do quantil da distribuição Normal *standard*, no entanto referem que resultados empíricos mostram que usar o quantil da T de *Student* é muito melhor do que usar o da Normal *standard*, quando n e m são pequenos.

Invertendo o teste, ie, resolvendo $|(\widehat{p}_1 - \widehat{p}_2) - (p_1 - p_2)| = k\widehat{\sigma}$, pode-se obter o correspondente intervalo de confiança. No entanto, considerando a definição, π deve ser menor do que $1 - (p_1 - p_2)m/(m+n)$ e maior do que $(p_1 - p_2)n/(m+n)$. Consequentemente, $\widehat{\pi}$ deve

satisfazer as mesmas condições (estas condições podem assegurar que $\widehat{\sigma} > 0$, mas o inverso não é verdadeiro). Estas considerações remeteram aqueles autores a tomarem o estimador truncado

$$\widetilde{p} = \begin{cases} (p_1 - p_2)/(m+n) & \text{se } \widehat{\pi} < (p_1 - p_2)n/(m+n) \\ \widehat{\pi} & \text{se } (p_1 - p_2)n/(m+n) \leq \widehat{\pi} \leq 1 - (p_1 - p_2)m/(m+n) \\ 1 - (p_1 - p_2)m/(m+n) & \text{se } \widehat{\pi} > 1 - (p_1 - p_2)m/(m+n) \end{cases}$$

em vez de $\widehat{\pi}$, e assim a estabelecerem o intervalo de confiança para $p_1 - p_2$ com a forma,

$$\left(\frac{\widehat{p}_1 - \widehat{p}_2}{1 + k^2/(m+n)} \pm \frac{k\sqrt{1 + k^2/(m+n)(1/m + 1/n)\widetilde{p}(1-\widetilde{p}) - (\widehat{p}_1 - \widehat{p}_2)^2/(m+n)}}{1 + k^2/(m+n)} \right).$$

Brown/Li denominaram o intervalo de confiança de recentrado, uma vez que é centrado por um valor $(1 + k^2/(m+n))^{-1}$. Apesar de parecer complexo, tem uma forma explicita e tem um comportamento muito bom.

Os referidos autores ([10]) apresentam uma comparação dos IC referidos para diferentes níveis de confiança ($\alpha = 0.01; 0.05$ e 0.1).

A título de resumo, apresentam-se algumas das suas conclusões:

- quando m e n são pequenos, o IC de Wald está abaixo do nível nominal;
- em muitos casos os IC Newcombe e recentrado são similares entre si e apresentam um bom comportamento;
- quando m e n são grandes, digamos $\min(m, n) \geq 50$, todos os IC têm um bom comportamento. No entanto, para o caso em que m e n são múltiplos e $p_1 - p_2$ é pequeno, o IC recentrado não é bom;
- muitos dos IC são muito conservativos quando p_1 e p_2 estão próximos de 0 ou 1;
- se $p_1 = p_2 = p$ a cobertura de todos os IC é simétrica em volta de $p = 0.5$;
- quando m e n são pequenos e $p_1 - p_2 \neq 0$ deve utilizar-se o IC de Newcombe ou o recentrado;
- quando $m \neq n$ e $p_1 - p_2 \approx 0$ deve-se utilizar o IC Newcombe.
- não deve utilizar-se o IC recentrado se m é um múltiplo de n ou vice-versa.

2.2 Distribuição Multinomial

2.2.1 Definição e Propriedades

A distribuição Multinomial é uma generalização da distribuição Binomial onde é permitida a divisão da população com mais do que duas simples categorias: "sucesso" ou "insucesso". A notação

$$\mathbf{Y}' = (Y_1, Y_2, \cdots, Y_t) \sim M_t(n, \mathbf{p} = (p_1, p_2, \cdots, p_t)')$$

indica que o vector $\mathbf{Y}$ segue uma distribuição Multinomial de dimensão t, baseada em n provas sendo $\mathbf{p}$ o vector de probabilidades de sucesso de cada uma das t categorias tal que $p_i \geq 0, i = 1(1)t$ e $\sum_{i=1}^{t} p_i = 1$. O vector $\mathbf{Y}$ tem domínio

$$\left\{ (y_1, y_2, \cdots, y_t) \in \mathbb{N}_0^t : \sum_{i=1}^{t} y_i = n \right\}.$$

A fmp da distribuição Multinomial é

$$P[\mathbf{Y} = \mathbf{y}] = P[Y_1 = y_1, Y_2 = y_2, \cdots, Y_t = y_t] = \frac{n!}{y_1! y_2! \cdots y_t!} p_1^{y_1} p_2^{y_2} \cdots p_t^{y_t}.$$

Sendo $E[\mathbf{Y}] = n\mathbf{p}$ e a sua matriz de covariância $\sum$ é $n(Diag(\mathbf{p}) - \mathbf{pp}')$, ou seja o elemento da linha i, coluna j da matriz $\sum$ é dado por

$$cov(Y_i, Y_j) = \begin{cases} np_i(1 - p_i) & , \text{se } i = j \\ -np_i p_j & , \text{se } i \neq j \end{cases}.$$

2.2.2 Estimação Pontual de p

Proposição 1. *Seja* $\mathbf{Y} \sim M_t(n, \mathbf{p})$*, o estimador de máxima verosimilhança de* $\mathbf{p}$ *é*

$$\widehat{\mathbf{p}} := \mathbf{Y}/n.$$

Demonstração. Para todo o $\mathbf{y} = (y_1, y_2, \cdots, y_t)$ pertencente ao domínio de $\mathbf{Y}$, a função verosimilhança de $\mathbf{p}$ é dada por

$$L(\mathbf{p}) = P[\mathbf{Y} = \mathbf{y}] = n! \prod_{i=1}^{t} \left(\frac{p_i^{y_i}}{y_i!} \right),$$

onde $p_i \geq 0$ tal que $\sum p_i = 1$.

Pretende-se mostrar que a função $L(\mathbf{p})$ é majorada por

$$n! \prod_{i=1}^{t} \left(\frac{y_i}{n} \right)^{y_i};$$

ou seja, que é suficiente que

$$\prod_{i=1}^{t} p_i^{y_i} \leq \prod_{i=1}^{t} \left(\frac{y_i}{n}\right)^{y_i}. \tag{2.8}$$

Observe-se que três situações podem ocorrer:

(i) existe um i tal que $p_i = 0$ e $y_i \geq 1$ ($\therefore y_i \neq 0$), ou

(ii) existe um i tal que $y_i = 0$ ou

(iii) $p_i > 0$ e $y_i > 0$ ($\forall i, i = 1(1)t$).

Se ocorre *(i)*, então $L(\mathbf{p}) = n! \prod_{i=1}^{t} p_i^{y_i} = 0 \leq n! \prod_{i=1}^{t} (\frac{y_i}{n})^{y_i}$ (condição universal).
Se ocorre *(ii)*, então, porque $0^0 = 1$ por continuidade, vem

$$L(\mathbf{p}) = n!(p_1^{y_1} \cdots p_i^{y_i} \cdots p_t^{y_t}) = n! \prod_{j=1, j\neq i}^{t} {p_j}^{y_j},$$

que corresponde ao caso *(iii)*.
Na situação *(iii)* pretende-se provar (2.8).
Recorde-se que a média geométrica e aritmética satisfazem

$$\left(\prod_{i=1}^{m} a_i\right)^{1/m} \leq \frac{1}{m} \sum_{i=1}^{m} a_i \quad \text{para } a_i > 0,\ i = 1(1)m \tag{2.9}$$

Aplicando (2.9), com $m = n = \sum_{i=1}^{t} y_i$ e $a' = \left(\frac{p_1}{y_1}, \cdots, \frac{p_1}{y_1}, \frac{p_2}{y_2}, \cdots, \frac{p_t}{y_t}, \cdots, \frac{p_t}{y_t}\right)$, onde há y_i cópias do factor p_i/y_i, para $i = 1(1)t$, resulta

$$\left(\left[\prod_{i=1}^{t} \left(\frac{p_i}{y_i}\right)^{y_i}\right]^{1/n}\right)^{n} \leq \left(\frac{1}{n} \sum_{i=1}^{t} y_i \frac{p_i}{y_i}\right) = \left(\frac{1}{n}\right)^{n}.$$

Consequentemente,

$$\prod_{i=1}^{t} \left(\frac{p_i}{y_i}\right)^{y_i} \leq \frac{1}{n^n},$$

e portanto,

$$\prod_{i=1}^{t} {p_i}^{y_i} \leq \frac{1}{n^n} \prod_{i=1}^{t} {y_i}^{y_i} = \prod_{i=1}^{t} \left(\frac{y_i}{n}\right)^{y_i}$$

e a prova fica concluída. □

Uma vez que se prova que a distribuição Multinomial pertence a uma família exponencial regular resulta que o estimador de máxima verosimilhança $\mathbf{p}$ tem as seguintes propriedades ([50]):

- é o estimador UMVUE de $\mathbf{p}$;
- é eficiente; ie, à medida que n aumenta indefinidamente para infinito, a matriz de covariância de $\sqrt{n}(\widehat{p}-p)$ aproxima-se do inverso da quantidade de informação de Fisher.

Na literatura são sugeridos outros estimadores alternativos para o MLE aplicando técnicas de Bayes e técnicas de *smoothing*, no entanto, não são considerados neste estudo.

2.2.3 Testes de Hipóteses sobre p

Aqui serão considerados os testes de hipótese nula simples *vs* hipótese alternativa global, para dados multinomiais.

Esta discussão requer uma revisão do princípio de construção de testes de hipóteses da razão de verosimilhança (LRT), *score* e Wald, para amostras de grandes dimensões.

Seja $\mathbf{Y} \sim M_t(n, \mathbf{p})$, com $\mathbf{p} = (p_1, p_2, \cdots, p_t)'$ desconhecido tal que $p_i \geq 0$ e $\sum p_i = 1$. Pretende-se testar

$$H_0 : \mathbf{p} = \mathbf{p}_0 \qquad vs \qquad H_1 : \mathbf{p} \neq \mathbf{p}_0,$$

onde $\mathbf{p}_0 = (p_1^0, p_2^0, \cdots, p_t^0)'$.

O teste **LRT** rejeita H_0 se

$$G^2 = 2\sum_{i=1}^{t} Y_i \ln\left(\frac{Y_i}{np_i^0}\right) \geq \chi^2_{\alpha,t-1}$$

onde $\chi^2_{\alpha,t-1}$ denota o quantil de ordem $1-\alpha$ da distribuição de Qui-quadrado com $t-1$ graus de liberdade (g.l.).

O teste ***score*** rejeita H_0 se

$$X^2 = \sum_{i=1}^{t} \frac{(Y_i - np_i^0)^2}{np_i^0} \geq \chi^2_{\alpha,t-1}$$

teste este que também se designa por teste do Qui-quadrado.

O teste de **Wald** rejeita H_0 se

$$W = \sum_{i=1}^{t} \frac{[Y_i - np_i^0]^2}{Y_i} \geq \chi^2_{\alpha,t-1}.$$

Este teste é também conhecido por teste do "Qui-quadrado mínimo modificado" para testar H_0.

Cressie e Read ([17]) apresentam um modo unificado de verificar estes testes introduzindo uma família de potência de divergência das estatísticas de teste denotada por $\{\mathbf{I}^\lambda : \lambda \in \mathbb{R}\}$,

que contem as estatísticas de teste G^2, X^2 e W como casos particulares.

A estatística $\mathbf{I}^\lambda$ é definida por

$$\mathbf{I}^\lambda := \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{t} Y_i \left\{ \left(\frac{Y_i}{n\widehat{p}_i^0} \right)^\lambda - 1 \right\}, \quad \text{para } \lambda \in \mathbb{R},$$

sendo I^0 e I^{-1} definidas por continuidade.

Observa-se que:

- $\mathbf{I}^0 = G^2$;
- $\mathbf{I}^1 = X^2$;
- $\mathbf{I}^{-2} = W$;
- $\mathbf{I}^{-1/2} = 4\sum_{i=1}^{t} \left[(Y_i)^{1/2} - (np_i^0)^{1/2} \right]^2$ denominada por estatística de Freeman-Turkey.

A questão que se coloca agora é qual destas estatísticas G^2, X^2, W ou outro qualquer membro da família apresentada por Cressie e Read, será melhor.

Alguns estudos feitos por estes autores, debruçaram-se sobre este tema e concluíram que a estatística $\mathbf{I}^\lambda$, para $\lambda \in [1/3, 3/2]$, e em particular para $X^2 = \mathbf{I}^1$, aproxima-se da distribuição de um Qui-quadrado com $t-1$ g.l. mais rapidamente do que as outras estatísticas $\mathbf{I}^\lambda$, para outros λ. No caso de X^2, um facto teórico que sugere a exactidão da distribuição limite χ^2_{t-1}, é a conformidade dos momentos das duas distribuições. Na realidade, quando $\mathbf{p} = \mathbf{p}^0$, para todo o n, verifica-se que

$$E[X^2] = \sum_{i=1}^{t} \frac{E[Y_i - np_i^0]}{np_i^0} = t - 1 = E[\chi^2_{t-1}].$$

Assim, os primeiros momentos de X^2 e χ^2_{t-1} são iguais. Similiarmente, mas acrescido de muitos cálculos, prova-se que

$$V(X^2) = 2(t-1)(1-1/n)$$

enquanto que $V(\chi^2_{t-1}) = 2(t-1)$. Assim, os segundos momentos concordam a menos de um factor $(1-1/n)$.

Cressie e Read recomendam usar $\mathbf{I}^{2/3}$ como estatística de referência.

Dado que as estatísticas X^2 e $\mathbf{I}^{2/3}$, ou alguma outra estatística $\mathbf{I}^\lambda$ com $\lambda \in [1/3, 3/2]$ são preferidas a G^2 e W, com base na rapidez com que se aproximam do valor nominal, algumas precauções devem ser tomadas por forma a que o ponto crítico do Qui-quadrado assimptótico possa ser usado sem problemas. Cochran ([14]) recomenda que se $np_i^0 \geq 5$, para todo o i,

então a aproximação para X^2 é adequada.

Estudos mais recentes mostram que se $np_i^0 \geq 1$, para todo o i com $np_i^0 \geq 5$ para 80% dos índices i, então a aproximação do Qui-quadrado é muito boa.

2.2.4 Estimação Intervalar para p

Existem várias formas de estabelecer intervalos de confiança para o parâmetro $\mathbf{p} = (p_1, p_2 \cdots, p_t)'$ e a diferença de proporções $p_i - p_j,\ i \neq j$ conhecidos.

Considere-se primeiro a construção de intervalos de confiança simultâneos para $p_i : 1 \leq i \leq t$. Denote-se $\mathbf{p}_{\_} := (p_1, \cdots, p_{t-1})'$ e $\widehat{\mathbf{p}}_{\_} := \left(\frac{Y_1}{n}, \cdots, \frac{Y_{t-1}}{n}\right)'$.

Gold ([26]) definiu a região de confiança para $\mathbf{p}$

$$R_G := \left\{ \mathbf{w} \in \mathbb{R}^{t-1} : (\widehat{\mathbf{p}}_{\_} - \mathbf{w})' \widehat{\Sigma}^{-1} (\widehat{\mathbf{p}}_{\_} - \mathbf{w}) \leq \frac{\chi^2_{\alpha,t-1}}{n} \right\}$$

onde $\widehat{\Sigma} = \text{Diag}\,(\widehat{\mathbf{p}}_{\_}) - \widehat{\mathbf{p}}_{\_}(\widehat{\mathbf{p}}_{\_})'$. Tendo observado que, $P_{\mathbf{p}}[\mathbf{p}_{\_} \in R_G]$ tende para $1 - \alpha$ à medida que n tende para infinito, uma vez que $\sqrt{n}(\widehat{\mathbf{p}}_{\_} - \mathbf{p}_{\_}) \to N_{t-1}(0, \text{Diag}(\mathbf{p}_{\_}) - \mathbf{p}_{\_}\mathbf{p}_{\_}')$

Gold propôs produzir como caso especial daquela região, intervalos de confiança simultâneos da forma

$$I_i^G = \left(\widehat{p}_i \pm \left(\frac{\widehat{p}_i(1-\widehat{p}_i)}{n} \right)^{1/2} (\chi^2_{\alpha,t-1})^{1/2} \right),\ i = 1(1)t.$$

Quesenberry e Hurst ([48]) propuseram uma região de confiança semelhante à de Gold e dada por

$$R_{QH} := R_{QH} := \left\{ \mathbf{w} \in \mathbb{R}^{t-1} : (\widehat{\mathbf{p}}_{\_} - \mathbf{w})' [\Sigma(\mathbf{w})]^{-1} (\widehat{\mathbf{p}}_{\_} - \mathbf{w}) \leq \frac{\chi^2_{\alpha,t-1}}{n} \right\}$$

onde $\Sigma(\mathbf{w}) = \text{Diag}(\mathbf{w}) - \mathbf{w}\mathbf{w}'$.

A região R_{QH} é assimptoticamente equivalente a R_G apenas diferem no facto de que R_G usa uma matriz de covariância estimada enquanto que $R_{_{QH}}$ usa a não estimada. R_G e R_{QH} são generalizações dos intervalos de confiança I_N e I_S para o parâmetro p da Binomial, respectivamente. O intervalo de confiança simultâneo de Quesenberry e Hurst para p_i é dado por

$$I_i^{QH} = \left(\frac{2Yi + c \pm \sqrt{c\left[c + \frac{4Yi(n-Yi)}{n}\right]}}{2(n+c)} \right) \tag{2.10}$$

onde $c = \chi^2_{\alpha,t-1}$.

A família de intervalos $\left\{I_i^G\right\}_{i=1}^t$ ou $\left\{I_i^{QH}\right\}_{i=1}^t$ não são satisfatórias. Os cálculos de Ghosh ([25]) mostraram que a distribuição usada nos intervalos $\left\{I_i^G\right\}_{i=1}^t$ pode não ser válida, e

que, mesmo que a distribuição assimptótica fosse exacta, ambas as famílias de intervalos são conservativas.

Goodman ([27]) estudou intervalos de confiança simultâneos derivando-os da desigualdade de Bonferroni. A desigualdade de Bonferroni diz que, para qualquer acontecimento E_j, tal que $P[E_j] \geq 1 - \alpha/t$, $j = 1(1)t$, se tem

$$P\left[\bigcap_{j=1}^{t} E_j\right] = 1 - P\left[\bigcup_{j=1}^{t} \overline{E_j}\right] \geq 1 - \sum_{j=1}^{t} P[\overline{E_j}] \geq 1 - \sum_{j=1}^{t} \alpha/t = 1 - \alpha.$$

Goodman aplica esta desigualdade de Bonferroni a acontecimentos

$$E_j := \left\{ \frac{n(\widehat{p}_j - p_j)^2}{p_j(1-p_j)} \leq \chi^2_{\alpha/t,1} \right\},\ j = 1(1)t,$$

para resultar na família de intervalos de confiança $\left\{I_i^{GM}\right\}_{i=1}^{t}$. Estes intervalos têm a mesma forma que os intervalos I_i^{QH} na equação (2.10) substituindo c por $\chi^2_{\alpha/t,1}$.

Em princípio, para casos de amostras de dimensão pequena pode-se utilizar individualmente os intervalos de confiança de Blyth/Still para cada p_i, com a constante de Bonferroni para obter intervalos de confiança simultâneos para $\{p_i\}_{i=1}^{t}$, com um nível de confiança $(1-\alpha)100\%$. No entanto, na prática isto não é fiável dado que apenas estão tabelados os intervalos de Blyth/Still para 95% e 99%. Uma alternativa computacional será usar a constante de Bonferroni com intervalos de cauda uma vez que eles podem ser construídos a partir de (2.5) para qualquer α.

Projecções da região de R_G de Gold podem ser usadas para obter intervalos de confiança simultâneos para a diferença de proporções, $p_i - p_j : 1 \leq i < j \leq t$. Os intervalos resultantes são dados por

$$\left(\widehat{p}_i - \widehat{p}_j \pm (\chi^2_{\alpha,t-1})^{1/2} \sqrt{\frac{\widehat{p}_i + \widehat{p}_j - (\widehat{p}_i - \widehat{p}_j)^2}{n}}\right) \tag{2.11}$$

tendo em conta que

$$\begin{aligned} V(\widehat{p}_i - \widehat{p}_j) = V(\widehat{p}_i) + V(\widehat{p}_j) - 2Cov(\widehat{p}_i, \widehat{p}_j) &= \frac{p_i(1-p_i)}{n} + \frac{p_j(1-p_j)}{n} - \frac{2(-p_i p_j)}{n} \\ &= \frac{p_i + p_j - (p_i - p_j)^2}{n}. \end{aligned}$$

Projecções baseadas na região de Quesenberry-Hurst, R_{QH}, parecem impossíveis de determinar uma vez que a variância de $(\widehat{p}_i - \widehat{p}_j)$ não depende de p_i nem de p_j, apenas de $p_i - p_j$. Goodman propôs usar intervalos de Bonferroni substituir em (2.11) o quantil $\chi^2_{\alpha,t-1}$ pelo quantil $\chi^2_{z,1}$, onde $z = \alpha/\binom{t}{2}$.

A literatura sugere que o sistema de intervalos de Goodman é o método de escolha de entre os referenciados.

2.3 Distribuição de Poisson

2.3.1 Definição e Propriedades

Diz-se que uma v.a. Y segue a distribuição de Poisson, com parâmetro $\lambda > 0$, e denota-se por $Y \sim P(\lambda)$, se a sua fmp é

$$P[Y = j] = \frac{e^{-\lambda}\lambda^j}{j!} \quad , \; j = 0, 1, ...$$

A média e a variância de Y são idênticas sendo $E[Y] = V[Y] = \lambda$.

A Figura 2.5 ilustra a forma da distribuição de Poisson através da fmp para diferentes valores do seu parâmetro.

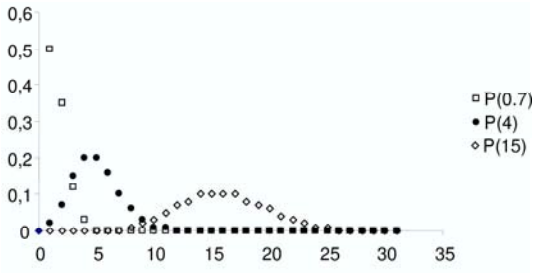


Figura 2.5: As fmp′s de variáveis com distribuição P(0.7), P(4) e P(15).

A distribuição de Poisson aparece como uma possível distribuição para a modelação de contagens de acontecimentos aleatórios no tempo ou no espaço sob suaves suposições (Karlin e Taylor, [31]) associados a sistemas de filas de espera. Surge também, como uma aproximação da distribuição Binomial, $B(n, p)$, quando n é grande e p é pequeno. Tal significa, que a distribuição de Poisson pode ser aplicada na modelação de contagens de acontecimentos raros num número elevado de provas.

Por exemplo, quando $Y \sim B(30, 0.1)$ tem-se

$$\begin{aligned} P[Y \leq 4] &= 0.825 \qquad \text{(valor exacto)} \\ &\approx 0.0.8153 \qquad \text{(valor aproximado)}, \end{aligned}$$

onde $W \sim P(\lambda = 30 \times 0.1)$, pelo que se comete um erro relativo de 1.1% na aproximação da Poisson à Binomial.

2.3.2 Gráfico de ajustamento de Hoaglin

Considere-se o problema de testar o ajustamento do modelo de Poisson baseado numa amostra aleatória $Y_1, ..., Y_t$ de dados de contagem. Hoaglin ([29]) propôs uma técnica gráfica análoga

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
F_k	57	203	383	525	532	408	273	139	45	27	10	4	0	1	1

Tabela 2.2: Numero de fluorescências a partir da Radioactividade do Polónio.

ao PP-plots para avaliar se esta amostra provêm de uma mesma população Y com distribuição $P(\lambda)$, para algum $\lambda > 0$. Essa técnica é baseada no facto de se tomar como estimativa de $P[Y = k]$ o valor de

$$\frac{F_k}{t} := \frac{\text{número de observações } Y_1, Y_2, \cdots, Y_t \text{ iguais a } k}{t}.$$

Assim, igualando o verdadeiro valor da probabilidade ao valor estimado vem

$$\frac{F_k}{t} = \frac{e^{-\lambda}\lambda^k}{k!}.$$

Aplicando logaritmo a ambos os membros, resulta

$$\ln F_k + \ln k! = \ln t - \lambda + k \ln \lambda,$$

donde, $\ln F_k + \ln k!$ é uma função linear de k. Assim, o ajustamento ao modelo de Poisson é adequado se os pontos pertencerem à linha recta com declive aproximadamente $\ln(\lambda)$ e ordenada na origem aproximadamente $(\ln(t) - \lambda)$. Nestes casos, um estimador de λ pode ser obtido a partir do gráfico e por dois modos possíveis: $\hat{\lambda} = e^{\text{declive da recta}}$ ou $\hat{\lambda} = (\ln t - \text{ordenada na origem})$. A não linearidade no gráfico indica desvios ao modelo de Poisson. Rutherford e Geiger (1910) apresentaram os dados da Tabela 2.2 relativos ao número de fluorescências devido à radioactividade do polónio em cada $t = 2608$ intervalos de tempo de $1/8$ minutos. Seja F_k o número de intervalos de tempo em que são observadas k fluorescências. Baseado nos valores da Tabela 2.3, procedeu-se à construção do gráfico de Hoaglin (ver Figura 2.6). Sobre estes dados, é visível quais são os pontos que não pertencem à recta de declive $\ln(\widehat{\lambda})$ e ordenada na origem $\ln(t) - \widehat{\lambda}$, onde $\widehat{\lambda} = 3.87$ é a estimativa de máxima verosimilhança de λ, assumindo que os dados resultam de uma amostra aleatória de uma distribuição de Poisson.

2.3.3 Teste de ajustamento

Uma das formas de averiguar o ajustamento dos dados a uma distribuição de Poisson consiste em usar a conhecida estatística de qui-quadrado de Pearson.

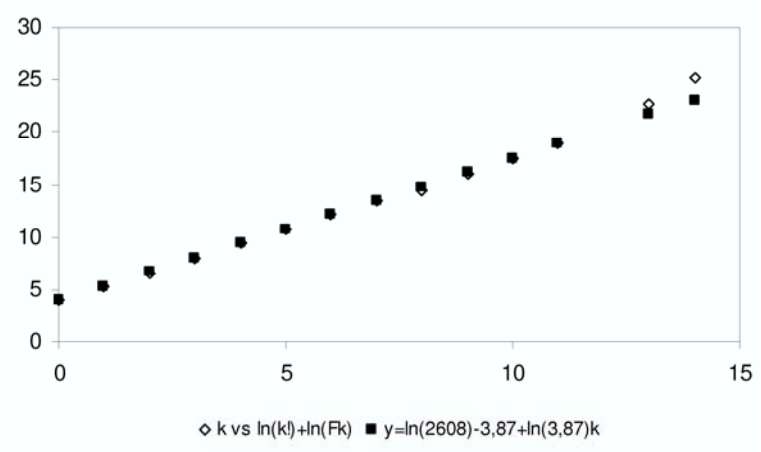


Figura 2.6: Gráfico de Hoaglin.

Genericamente, considerando uma população X dividida em m classes $A_1, A_2, \cdots, A_m$, mutuamente exclusivas, o teste sugerido por Karl Pearson para testar

$$H_0 : P[X \in A_j] = p_{0j},\ j = 1, 2, \cdots, m, \tag{2.12}$$

onde $p_{0j} \in (0, 1)$, $\sum p_{0j} = 1$, consiste em considerar a estatística de teste

$$\chi^2 = \sum_{j=1}^{m} \frac{(N_j - np_{0j})^2}{np_{0j}} \tag{2.13}$$

onde N_j é a v.a. que representa a frequência observada na amostra da classe A_j e n o tamanho da amostra.

A estatística de Pearson χ^2 é, na realidade, uma medida do afastamento entre a frequência observada e a frequência esperada sob a validade da hipótese (2.12). Evidentemente, quanto menor for o valor observado para χ^2, mais plausível é a hipótese.

A v.a. χ^2, definida por (2.13), tem assimptoticamente (quando $n \to \infty$) distribuição de Qui-quadrado com $(m - 1)$ g.l..

Na prática, caso em que o teste é usado com n finito, é usual trabalhar-se com classes A_j cujas frequências esperadas, sob H_0, np_{0j}, sejam não inferior a cinco, de modo que a aproximação da estatística χ^2 à sua distribuição limite seja "mais" válida ([39]).

No caso particular de averiguar o ajustamento dos dados a uma população $X \sim P(\lambda)$, com $\lambda > 0$ conhecido, tem-se que $p_{0j} = P[X \in A_j | X \sim P(\lambda)]$. Se λ é desconhecido, p_{0j} deve ser estimado através do método da máxima verosimilhança. E, nesse caso, prova-se que, a estatística de Pearson, χ^2, terá uma distribuição assimptótica de um qui-quadrado com $(m - k - 1)$ g.l., quando o tamanho da amostra aumenta indefinidamente.

Neyman-Pearson estabeleceram regiões críticas para o teste de ajustamento da estatística de Pearson. Se, para uma amostra concreta, se contarem n_j observações na classe $A_j, j = 1, 2, \cdots, m$, com $\sum n_j = n$, tem-se

$$\chi^2_{obs} = \sum_{j=1}^{m} \frac{(n_j - np_{oj})^2}{np_{0j}}.$$

Valores grandes de χ^2_{obs} traduzem o afastamento dos dados em relação à hipótese

$$H_0 : X \sim P(\lambda).$$

De forma a construir uma região crítica, com nível de significância α, procura-se o quantil de ordem $1 - \alpha$ da distribuição de qui-quadrado com $m - k - 1$ g.l. e rejeita-se a hipótese (2.12) se $\chi^2_{obs} \geq \chi^2_{\alpha, m-k-1}$.

Aplicando este teste aos dados apresentados na Tabela 2.2, obteve-se $\chi^2_{obs} = 12,8$ (ver Tabela 2.3). Neste caso $m = 11$ e $k = 1$. Para $\alpha = 0,05$ e para $m-k-1 = 9$ g.l. tem-se, $\chi^2_{0.05} = 16,92$. Como $\chi^2_{obs} < 16,92$, a decisão é de não rejeitar, ou seja, os dados são compatíveis com a hipótese proposta de que os dados provêm de uma distribuição de Poisson.

k	F_k	$ln(k!) + ln(F_k)$	Frequências Observadas	Frequências Esperadas	Estatística Teste
0	57	4.043051	57	54.314424	0.132788
1	203	5.313206	203	210.280961	0.252103
2	383	6.641182	383	407.056531	1.421711
3	525	8.055157	525	525.313113	0.000187
4	532	9.454697	532	508.443876	1.091352
5	408	10.798759	408	393.693083	0.519917
6	273	12.188723	273	254.033682	1.416037
7	139	13.459635	139	140.500553	0.016026
8	45	14.411265	45	67.994348	7.776235
9	27	16.097664	27	29.249273	0.172969
10	10	17.406997	16	15.798428	0.002572
11	4	18.888602			
13	1	22.552163			
14	1	25.191221			
Total	2608			2606.678	12.8019

Tabela 2.3: Cálculos auxiliares para a construção do gráfico de Hoaglin e aplicação do teste de Pearson.

Capítulo 3

Testes Simultâneos em Tabelas de Contingência

3.1 Introdução

Um dos procedimentos mais comuns ao nível de testes de hipóteses em tabelas de contingência de dupla entrada é o teste de independência usando a estatística de qui-quadrado de Pearson. A estatística de Pearson avalia uma *distância* entre o que se observa e o que seria esperado, sob a validade da hipótese de independência. A ideia base na regra de decisão deste teste é a seguinte: se o valor observado daquela distância é *pequeno*, do ponto de vista estatístico, não deve ser rejeitada a hipótese de independência entre as duas variáveis categorizadas representadas, uma em linha e outra em coluna, na tabela; se aquele valor é *grande*, então a suposição de independência deve ser rejeitada.

Contudo, a estatística de Pearson, só por si, não identifica a independência em células individuais da tabela de contingência. Na realidade, este teste é "global" pois não estabelece a significância da associação em cada célula da tabela mas na tabela no seu global.

O problema da identificação da existência ou não de associação entre categorias numa tabela de contingência reveste-se de especial importância. Lancaster ([56]) mostrou que qualquer tabela $r \times c$ pode ser particionada em $(r-1) \times (c-1)$ tabelas independentes de dimensão 2×2. A interpretação de tabelas de dimensões reduzidas é simples, no entanto, a aplicação deste método de partição para tabelas de contingência grandes, torna-se muito complicado dado que gera muitas tabelas de dimensão 2×2. Por exemplo, uma tabela de 10×10 produz 81 tabelas de dimensão 2×2, o que faz com que a extracção de informação seja difícil.

Haberman ([28]) estabeleceu uma análise de cada uma das parcelas da estatística de qui-

quadrado de Pearson definindo uma estatística residual standardizada e ajustada (STAR) para cada célula, tendo mostrado que STAR segue assimptoticamente uma distribuição Normal *standard* sob a validade da hipótese de independência. Assim, calculando, para cada célula da tabela, o valor observado para a estatística STAR, pode estabelecer-se se este valor é significativo. Um valor maior do que o quantil de ordem $1-\alpha$ de uma distribuição $N(0,1)$ indicará falta de ajustamento da estatística STAR à distribuição Normal *standard*, ao nível de significância α. Por outras palavras a célula correspondente estará associada a categorias, das variáveis linha e coluna, significativamente responsáveis pela rejeição da hipótese global de independência. Este método, de usar a estatística STAR é simples mas, sob considerações simultâneas de todas as células na tabela de contingência, produz muitos falsos positivos ([55]).

Outro método foi também introduzido por Haberman recorrendo ao gráfico de probabilidade da Normal para os valores da estatística STAR. No entanto, a interpretação deste gráfico é frequentemente subjectiva, particularmente quando o número de células a ser testado é grande.

Há, assim, a necessidade de estabelecer outros procedimentos que testem a independência em cada célula para tabelas de contingência (de dimensão grande ou não).

Kim e Tsui ([32]) analisaram formas de testar a independência das categorias em células individuais de uma tabela de contingência baseado em testes simultâneos.

Em problemas de testes simultâneos, nomeadamente quando a hipótese de teste H_0 pode ser escrita como uma intersecção de m hipóteses nulas $H_{0,i}$, a taxa de erro na decisão de rejeitar H_0 é controlada tendo em conta as taxas de erro individuais associadas às hipóteses $H_{0,i}$. Vários autores ([4], [55] e [56]) observaram que, por exemplo, quando $H_{0,i}$ são hipóteses simples, a probabilidade de que, no mínimo, um dos testes i conduza à rejeição de H_0 quando se deveria aceitar H_0, cresce exponencialmente com o número m de hipóteses.

Na secção seguinte serão apresentadas três formas de definir taxas de erro global associadas à regra de decisão em testes simultâneos e vários procedimentos que permitem controlar um majorante para essas taxas de erro. Estes resultados derivam, em particular, dos trabalhos de [4], [32], [54], [55] e [56].

	Aceitar H_0	Rejeitar H_0	Total
H_0 verdadeira	U	V	m_0
H_0 falsa	T	S	m_1
Total	W	R	m

Tabela 3.1: Possíveis resultados de um m-teste de hipóteses.

3.2 Procedimentos de controlo em testes simultâneos

Considere-se um teste de hipóteses onde a hipótese nula é definida pela intersecção de m hipóteses nulas simples, ou seja, $H_0 := \bigcap_{i=1}^{m} H_{0,i}$.

A Tabela 3.1 descreve os variados resultados que se podem obter quando se aplicam m testes de hipóteses simultaneamente para testar H_0. O número m de hipóteses é conhecido, mas as quantidades m_0 e m_1 de hipóteses nulas verdadeiras e falsas são parâmetros desconhecidos. O número de hipóteses nulas rejeitadas, R, é observável. Enquanto que o número de falsos positivos, V, o número de falsos negativos, o número de verdadeiros negativos, e o número de verdadeiros positivos, são variáveis aleatórias não observáveis.

3.2.1 Family-Wise Error Rate (FWER)

Uma das taxas de erro classicamente usadas na literatura, conhecida por FWER, é definida como a probabilidade de gerar uma ou mais falsas rejeições, ie,

$$FWER = P[V \geq 1] = P[\text{Rejeitar } H_{0,i},\ i = 1(1)m | H_0 \text{ Verdadeira}],$$

onde V é o número de hipóteses rejeitadas quando a hipótese é verdadeira.

Existe uma variedade de métodos que controlam a FWER, sendo o mais usado o método de Bonferroni. Este método rejeita H_0, a um nível se significância α, se para algum i, a hipótese $H_{0,i}$ é rejeitada a um nível de significância $\alpha_i = \alpha/m$, ou seja, se $p_i \leq \alpha_i$, onde p_i é o $p-value$ associado ao teste de hipóteses $H_{0,i}$. Portanto, a FWER total é menor ou igual do que α.

Outros métodos Family-Wise foram desenvolvidos para melhorar a potência do método de Bonferroni, mas eles dificilmente podem rejeitar a hipótese nula quando é realmente falsa. Em particular, a proporção de falsas hipóteses nulas que são correctamente rejeitadas decresce significativamente à medida que m aumenta.

3.2.2 False Discovery Rate (FDR)

Bejamini e Hochberg ([4]) introduziram uma nova taxa de erro, denotada por FDR, e definida como sendo a proporção esperada de falsos positivos de entre todas as hipóteses nulas rejeitadas, ou seja,

$$E\left[\frac{V}{R} \mid R > 0\right] P[R > 0].$$

Diversas propriedades importantes da FDR foram discutidas em [4] e [59]. Uma das vantagens da FDR reside na identificação do maior número de hipóteses significativas enquanto mantém um pequeno número de falsos positivos. Mas quando $m_0 = m$, a FDR é equivalente à FWER. Bejamini e Hochberg apresentam um método de $p - values$ ordenados que controla a FDR, garantindo que FDR$\leq \alpha$, assumindo que

$$Y_1, Y_2, ..., Y_m$$

são m estatísticas de teste independentes para testar, respectivamente, $H_{0,1}$, $H_{0,2}$, $\cdots$, $H_{0,m}$ com $p - values$ correspondentes

$$P_1, P_2, ..., P_m,$$

e $p - values$ ordenados

$$P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}.$$

O método é definido pelo seguinte algoritmo,

1. Fixe-se α, onde $0 \leq \alpha \leq 1$;
2. Calcule-se $\widehat{i} = \max\left[i : P_{(i)} \leq \frac{i}{m}\frac{\alpha}{\pi_0}\right]$, onde $\pi_0 (= m_0/m)$ denota a proporção de $H_{0,i}$ verdadeiras;
3. Se $\widehat{i} \geq 1$, rejeitam-se todas as hipóteses $H_{0,i}$ com $P_i \leq P_{(i)}$;
   Se $\widehat{i} = 0$, não se rejeita nenhuma hipótese $H_{0,i}$; portanto, não se rejeita H_0.

Em geral, $\pi_0 = 1$ é a escolha mais conservativa possível. Diversos estudos discutem estimativas para π_0 ([59] e [21]).

3.2.3 Positive False Discovery Rate (pFDR)

Storey ([55] e [56]) introduziu a pFDR e definida por

$$pFDR = E\left[\frac{V}{R} \mid R > 0\right].$$

O termo "positive" foi incluído porque se assume que no mínimo uma hipótese significativa deve ocorrer.

Storey apresenta uma forma para controlar a pFDR em tudo semelhante ao procedimento de Bejamini-Hochberg, diferindo no facto de que Storey estima o valor de π_0, tendo mostrado que os dois procedimentos são equivalentes quando $\widehat{\pi}_0 = 1$. No entanto, se $\widehat{\pi}_0 < 1$ e $\widehat{\pi}_0$ é correctamente estimado, o procedimento de Storey é mais potente quando os dois procedimentos controlam a mesma taxa FDR, ou seja, os procedimentos de Storey e Bejamini-Hochberg rejeitam o mesmo número de hipóteses, no entanto o procedimento de Storey tem uma menor FDR ([55]).

A proporção estimada das hipóteses $H_{0,i}$ verdadeiras, $(\widehat{\pi}_0)$, pode ser obtida a partir de *software* desenvolvido por Storey e Dabney (http://faculty.washington.edu/jstorey).

A seguir descreve-se o procedimento de controlo de Storey:

1. Fixe-se α, onde $0 \leq \alpha \leq 1$;
2. Estime-se $\widehat{\pi}_0$;
3. Calcule-se $\widehat{i} = \max\left[i : \widehat{pFDR}_\lambda(P_{(i)}) \leq \alpha\right] = \max\left[i : P_{(i)} \leq \frac{i}{m}\frac{\alpha}{\widehat{\pi}_0}\right]$;
4. Se $\widehat{i} \geq 1$, rejeitam-se todas as hipóteses $H_{0,i}$ com $P_i \leq P_{(i)}$;
   Se $\widehat{i} = 0$, não se rejeita nenhuma hipótese $H_{0,i}$; portanto, não se rejeita H_0.

onde λ é um parâmetro que condiciona a estimativa de π_0, ou seja $\widehat{\pi}_0$. O programa implementado no *software* R para o cálculo de $\widehat{\pi}_0$ permite a escolha quer do valor de λ quer de métodos para estimar π_0, no entanto, se não se seleccionar nenhuma opção, por omissão é considerado o método *smoother* proposto por Storey ([56]), que é considerada a melhor opção. Se, por outro lado se considerar $\lambda = 0$, então obtém-se a estimativa implícita na metodologia de Bejamini-Hochberg ([4]).

Quando se realiza um teste de hipóteses para testar uma hipótese nula genérica H_0, é medido o nível de erro cometido na decisão do teste. O $p - value$ é uma medida da significância do erro em termos da razão de falsos positivos. Por exemplo, se $p - value = 0.05$ indica que 5% é a percentagem de erro mínimo cometido quando a estatística de teste conduz à rejeição de H_0, sendo esta verdadeira. No entanto, em testes simultâneos o $p - value$ não fornecerá uma medida dos erros entre as hipóteses nulas declaradas de significativas. Esta informação pode ser fornecida pela pFDR, que dá uma medida da proporção de falsos positivos entre todas as hipóteses significativas.

	1	2	$\cdots$	c	
1	N_{11}	N_{12}	$\cdots$	N_{1c}	N_{1*}
2	N_{21}	N_{22}	$\cdots$	N_{2c}	N_{2*}
$\cdots$			$\cdots$		$\cdots$
r	N_{r1}	N_{r2}	$\cdots$	N_{rc}	N_{r*}
	N_{*1}	N_{*2}	$\cdots$	N_{*c}	N_{**}

Tabela 3.2: Tabela de contingência $r \times c$.

Storey definiu uma quantidade análoga ao $p-value$, mas em termos da pFDR, que denominou por $q - value$. O $q - value$ pode ser visto como uma proporção esperada de falsos positivos e representa a menor taxa pFDR que pode ocorrer quando a estatística de teste conduz à rejeição de H_0 num dado conjunto de regiões de rejeição. Por exemplo, se $q - value = 0.05$ indica que 5% é a percentagem de erro cometido de uma hipótese significativa ser declarada falsamente como positiva.

Denotando por R_α a região de rejeição de H_0, a um nível de significância α, definida em termos de uma estatística de teste T, tem-se que

$$p - value = \inf_{\{R_\alpha:\ t \in R_\alpha\}} P[T \in R_\alpha \mid H_0 \text{ é verdadeira}],$$

e

$$q - value = \inf_{\{R_\alpha:\ t \in R_\alpha\}} P[H_0 \text{ é verdadeira} \mid T \in R_\alpha] = \inf_{\{R_\alpha:\ t \in R_\alpha\}} pFDR(R_\alpha).$$

Em testes simultâneos o procedimento de Bonferroni controla a FWER através do cálculo dos $p - values$ associados a cada uma das hipóteses nulas $H_{0,i}$.

O procedimento de Storey controla a pFDR usando os $q - values$ associados a cada uma das hipóteses nulas.

3.3 Procedimento proposto para testes simultâneos numa tabela de contingência

Uma tabela de contingência de dupla entrada é apresentada na Tabela 3.2. Para testar a independência entre as duas variáveis, uma em linha (com r categorias) e outra em coluna (com c categorias), ie,

$$H_0 : p_{ij} = p_{i*} \cdot p_{*j}, \quad i = 1, 2, \cdots, r, \quad j = 1, 2, \cdots, c, \tag{3.1}$$

usualmente recorre-se à estatística de Pearson (3.2)

$$\chi^2 = \sum_i \sum_j \frac{(N_{ij} - E_{ij})^2}{E_{ij}}, \tag{3.2}$$

onde N_{ij} e E_{ij} são os valores observados e esperados na célula correspondente à i-ésima linha e j-ésima coluna da tabela.

Testar a hipótese nula (3.1) corresponde a testar múltiplas hipóteses nulas $H_{0,i}$ simultaneamente. Mais precisamente, as $r \times c$ hipóteses nulas

$$H_{0,1\times 1} : p_{11} = p_{1*} \cdot p_{*1}$$

$$H_{0,1\times 2} : p_{12} = p_{1*} \cdot p_{*2}$$

$$\dots$$

$$H_{0,r\times c} : p_{rc} = p_{r*} \cdot p_{*c}$$

Para averiguar se uma célula da tabela de contingência é responsável pela eventual rejeição da hipótese nula de independência, (3.1), pode-se recorrer ao cálculo dos resíduos ajustados (em inglês, *Standardized Adjusted Residual* - STAR), definidos por

$$e_{ij} = \frac{N_{ij} - \frac{N_{i*}N_{*j}}{N_{**}}}{\left(\frac{\nu_{ij}N_{i*}N_{*j}}{N_{**}}\right)^{1/2}} = \frac{N_{ij} - E_{ij}}{(\nu_{ij}E_{ij})^{1/2}},$$

onde

$$\nu_{ij} = \left(1 - \frac{N_{i*}}{N_{**}}\right)\left(1 - \frac{N_{*j}}{N_{**}}\right).$$

Sob a hipótese nula de independência (3.1), Haberman ([28]) provou que

$$e_{ij} \to N(0,1), \text{ quando } N_{**} \to \infty$$

A completa derivação pode ser encontrada em Haberman([28]) e Agresti ([2]). Agresti ([2]) mencionou que o valor absoluto de e_{ij}, que excede em 2 (ou em alguns casos em 3), indica uma associação significativa entre a categoria i da variável linha e a categoria j da variável coluna. Mas tal significância é medida individualmente e não tem em conta a significância global do teste (3.1).

Assim, torna-se pertinente a aplicação de procedimentos para testes simultâneos com o objectivo de identificar as células significativas da tabela de contingência controlando a significância global da tabela. Propõe-se então aplicar os procedimentos de testes simultâneos descritos para identificar as células significantes na tabela de contingência, usando a estatística

STAR, e_{ij}, para testar $H_{0,i\times j}$.

Com base no estudo de simulação, Kim e Tsui ([32]) apresentam uma comparação entre os três procedimentos e o teste individual tendo concluído que:

- a potência de todos os procedimentos cresce quando a dimensão da amostra aumenta;
- se o número de hipóteses significativas é grande, o procedimento de Storey produz grande potência com pequeno erro do tipo I comparado com o procedimento do teste individual;
- a potência dos procedimentos é ordenada da seguinte forma:

$$\text{Storey} > \text{Benjamini-Hochberg} > \text{Bonferroni}.$$

Capítulo 4

Análise estatística de dados genómicos

4.1 Introdução

Nos últimos anos tem-se vindo a assistir a um avanço extraordinário da Genética com a descodificação do código genético completo de um número cada vez maior de espécies. Com base nesses dados, um dos desafios colocados à comunidade científica é extrair informação estatística relevante que permita identificar propriedades associadas às sequências de símbolos genéticos (nucleótidos, codões ou aminoácidos).

O grupo de Bioinformática da Universidade de Aveiro tem estado a desenvolver um sistema informático (denominado Anaconda) que fornece um conjunto de ferramentas estatísticas, bioinformáticas e de visualização de dados para análise primária da estrutura de genes. Em [47] são apresentadas algumas das ferramentas estatísticas implementadas.

Neste estudo são consideradas sequências de codões (tripletos de símbolos genéticos) com início no codão de iniciação (AUG) e fim num dos codões stop (UAA, UAG e UGA), nessas sequências combinam-se codões num texto sem espaços. No Apêndice A exemplifica-se um gene sequenciado da espécie *H.sapiens*.

As sequências completas de genomas amostrais já descodificadas e acessíveis através de bases públicas, como sejam por exemplo o Genbank (ftp://ftp.ncbi.nih.gov/genbank/gennomes), são descarregadas automaticamente pelo Anaconda. Em cada sequência, o Anaconda, imitando o ribossoma na tradução do mRNA, identifica o codão de iniciação e inicia a leitura, de três em três nucleótidos, na direcção $3'$ até encontrar um dos três codões stop, ([38]). Por cada vez que é lido um codão, o Anaconda memoriza os codões imediatamente antes

e após, vizinhos na direcção $3'$ e $5'$: A - site codon e E - site codon, respectivamente (ver Figura 1.7). Através da leitura de todas as sequências completas das zonas codificantes do genoma de qualquer espécie, o Anaconda constrói a tabela de contingência das frequências absolutas de todos os pares de codões justapostos existentes nesse genoma. Existem $4^3 = 64$ codões distintos sendo 3 codões stop; logo, representando em linha todos os possíveis codões que podem aparecer como primeira componente do par de codões consecutivos, e em coluna todos os possíveis codões que podem aparecer como segunda componente, aquela tabela de contingência terá dimensão 61×64. Na Figura 4.1 encontra-se um excerto da tabela de contingência relativa à espécie *Homo Sapiens*. Assim, por exemplo, ao longo de todo o código genético da espécie considerada, o par de codões AAA-AAA ocorre 6643 vezes.

	3' AAA	3' AAC	3' AAG	3' AAU	3' ACA	3' ACC	3' ACG	3' ACU	3' AGA
Lys AAA	6643	5066	9307	6125	6193	5039	1333	4834	5420
Asn AAC	6575	6257	8628	4046	5415	5424	1860	3590	4730
Lys AAG	15080	8141	19355	6641	6072	6544	2562	4400	6650
Asn AAU	5341	2645	3733	3150	2587	1912	553	1988	1943
Thr ACA	4340	2810	4791	3220	3392	2525	759	2698	2824
Thr ACC	5919	6005	9995	3691	5472	8011	2577	4024	3318
Thr ACG	896	694	1285	532	813	975	425	682	621
Thr ACU	2643	1449	2046	1612	2064	1849	443	1676	1358
Arg AGA	6267	3799	5951	4523	3426	2609	821	2915	4828
Ser AGC	6591	5795	8173	3682	5373	7222	2066	4113	4762
Arg AGG	7226	4185	9083	3516	3944	3642	1464	2957	4154
Ser AGU	3235	1912	2111	2068	1837	1601	386	1559	1364
Ile AUA	4104	1864	3144	2547	1990	1159	383	1643	1621
Ile AUC	5598	6196	8190	4363	4746	7261	1780	4222	3372
Met AUG	7004	5409	10140	5149	4299	5038	1435	4047	4088
Ile AUU	4284	1978	3188	2761	2097	1674	408	1957	1494
Gln CAA	4667	3106	4791	3901	3484	2366	861	2522	3653
His CAC	4419	4325	6543	2892	5007	5297	3012	4816	4474
Gln CAG	11157	8054	14885	6568	6551	7190	2866	4587	7915
His CAU	2864	1289	2665	1696	1678	1405	459	2763	1280
Pro CCA	4032	2986	5407	3780	3481	2980	849	2716	3480
Pro CCC	6299	6645	10819	3440	6222	7914	3607	4647	4858
Pro CCG	943	756	1669	599	852	1232	588	625	959
Pro CCU	3073	1681	2930	1951	2370	2106	667	2133	1617
Arg CGA	1469	1022	1944	1192	1061	983	283	686	1371
Arg CGC	1823	2057	4062	942	1923	2814	1515	1194	1734
Arg CGG	2706	2165	4645	1348	1496	2397	931	1224	1922
Arg CGU	770	476	749	438	452	563	155	495	490
Leu CUA	3009	1697	3357	2259	1755	1270	495	1545	1738
Leu CUC	4563	6667	8764	4081	5099	8274	2450	4687	3792

Figura 4.1: Extracto da tabela de frequências absolutas da espécie *Homo Sapiens*.

Com base em metodologias estatísticas para dados categorizados organizados em tabelas de contingência, o Anaconda testa automaticamente a hipótese de não existência de associação entre codões justapostos usando a estatística de qui-quadrado de Pearson e constrói o chamado mapa de contexto dos pares codões ([38]). A estatística de qui-quadrado de Pearson é dada por 3.2. O mapa de contexto dos pares codões não é mais do que a matriz dos valores STAR, $[e_{ij}]_{i,j}$, segundo a fórmula 3.3, discretizados numa escala de cores. A Figura

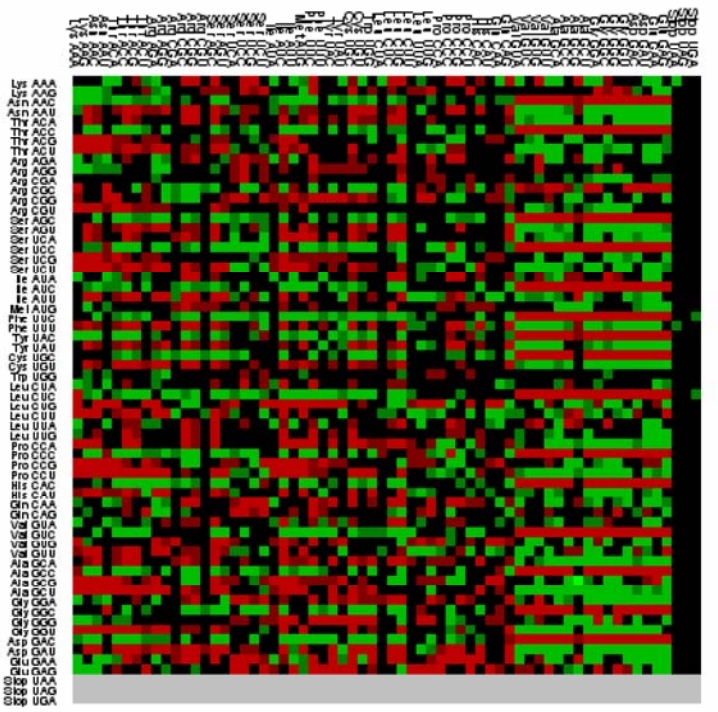


Figura 4.2: Mapa Contexto $3'$ - Imagem correspondente à matriz dos resíduos da espécie *Homo Sapiens*.

4.2 ilustra o mapa de contexto dos pares de codões para a espécie *Homo Sapiens*. A escala de cor varia de vermelho a verde passando pelo preto de acordo com a ordem de grandeza dos valores STAR. Os valores STAR quantificam a preferência de justaposição de codões face à independência, sendo que valores *significativamente* positivos estão coloridos a verde e estão associados a pares de codões **preferidos** no genoma, os valores *significativamente* negativos estão coloridos a vermelho e estão associados aos que são **preteridos**, face à independência e os restantes valores estão coloridos a preto e estão associados a pares de codões para os quais a independência não é rejeitada. Na Figura 4.2 pode ser observado que, por exemplo, o par AAG-AAG mostra maior preferência de ocorrência, face à não existência de associação entre codões consecutivos.

4.2 Procedimentos para definir significância de pares de codões

Nesta secção aplicam-se os procedimentos expostos no Capítulo 3, para definir os pares de codões significantes. Consideram-se os genomas das espécies *Saccharomyces cerevisiae*, *Can-*

dida albicans, *Thermoplasma acidophilum* e *Bacillus cereus*. Ilustrando os mapas de contextos dessas espécies resultantes de todos os procedimentos aplicados.

As espécies *Saccharomyces cerevisiae* e *Candida albicans*, pertencentes ao domínio *Eukarya*, são duas espécies de referência muito consideradas em estudos da Biologia Molecular, pensa-se que têm o mesmo ancestral, em que a segunda terá degenerado e a primeira mantido as características genéticas do ancestral. Para estas espécies expõe-se aqui detalhadamente a metodologia adoptada. Nos Apêndices C e D, respectivamente, ilustram-se as figuras e tabelas resultantes da aplicação de todas as metodologias para as espécies *Thermoplasma acidophilum* e *Bacillus cereus* como representantes dos domínios *Archaea* e *Bacteria*, respectivamente.

Todos os cálculos e resultados finais podem ser consultados em D:\Análise estatística de dados genómicos\Alínea 4.2.

No caso concreto as tabelas de contingência a estudar são de 61×64 e assim o número possível de pares são $3904 (= 61 \times 64)$.

Para estas 4 espécies, a tarefa é identificar, no contexto de testes simultâneos, quais daqueles pares estão significativamente associados. Mais precisamente, o objectivo é testar a hipótese de independência (global) entre pares de codões consecutivos a qual é equivalente à intersecção de 3904 hipóteses:

$$H_{0,1\times 1} : p_{11} = p_{1*} \cdot p_{*1}$$

$$H_{0,1\times 2} : p_{12} = p_{1*} \cdot p_{*2}$$

$$\dots$$

$$H_{0,61\times 64} : p_{61\times 64} = p_{61*} \cdot p_{*64},$$

seguindo a notação introduzida no Capítulo 3.

Aplicando os procedimentos para controlo das taxas de erro em testes simultâneos: Bonferroni (BF), Bejamini-Hochberg (B-H) e Storey (ST), em confronto com os resultados dados pela estatística STAR (IND), obtiveram-se os resultados para as espécies *Saccharomyces cerevisiae* e *Candida Albicans* que se resumem nas Tabelas 4.1 e 4.3, respectivamente. Os pares de codões com valores da estatística STAR mais extremos são obviamente dados como significativos para todos os procedimentos. O valor crítico para o qual um par de codão deixa de ser significativo varia de procedimento para procedimento.

Nas Tabelas 4.2 e 4.4 sumariam-se as contagens de pares significantes (preferidos e preteridos) por cada procedimento.

i	Par codão	STAR	p-value	IND	BF	B-H	ST
1	CAG - CAG	45.422	0	S	S	S	S
2	GCU - GCU	36.231	0	S	S	S	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
668	CAG - CCG	4.719	0.000002	S	S	S	S
669	GAC - AAG	4.695	0.000002	S	N	S	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
964	UUA - UAA	3.029	0.002453	S	N	S	S
965	CGA - UUU	3.026	0.002478	S	N	N	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1084	CUU - GUG	2.579	0.009908	S	N	N	S
1085	GAC - CCU	2.575	0.010024	N	N	N	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1103	CCU - CGC	2.518	0.011802	N	N	N	S
1104	CUC - AGU	2.499	0.012454	N	N	N	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2577	AGA - GAU	-2.184	0.028962	N	N	N	N
2578	CGA - GCG	-2.19	0.028524	N	N	N	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2690	CCC - UGC	-2.572	0.010111	N	N	N	S
2691	AUU - CGC	-2.577	0.009966	S	N	N	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2725	CAA - GAU	-2.697	0.006996	S	N	N	S
2726	GUC - UCU	-2.707	0.006789	S	N	S	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3229	CGA - GCU	-4.689	0.000001	S	N	S	S
3230	AGU - AUG	-4.712	0.000002	S	S	S	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3904	UUU - AAG	-24.58	0	S	S	S	S

Tabela 4.1: Pares de codões significativos para a espécie *Saccharomyces cerevisiae*, segundo os quatro procedimentos: Individual(IND), Bonferroni (BF), Bejamini-Hochberg (B-H) e Storey (ST)) para controlo da taxa de erro correspondente ($p - value$, FWER, FDR e pFDR) com $\alpha = 0.01$. S e N denotam pares significantes e pares não significantes respectivamente.

Procedimento	#pares sig. preferidos	#pares sig. não preteridos
Teste Individual	1084	1214
Bonferroni	668	675
Bejamini-Hochberg	964	1179
Storey	1103	1327

Tabela 4.2: Resumo dos resultados dos testes simultâneos - da espécie *Saccharomyces cerevisiae*.

Para ambas as espécies, comparando os procedimentos que controlam a FWER$\geq$ 0.01, o teste individual encontrou mais pares significativos do que o procedimento de Bonferroni. De entre os procedimentos que controlam a FDR$\geq$ 0.01, o procedimento de Storey encontrou um maior número de pares de codões significantes do que o procedimento de Bejamini-Hochberg. Por exemplo, para a *Saccharomyces cerevisiae*, o teste individual encontrou 2298 pares (preferidos e preteridos) e o teste de Bonferroni encontrou 1343 pares ambos significativos com probabilidade mínima de encontrar um falso positivo igual a 0.01. Os outros métodos, que controlam a FDR podem ser interpretados similarmente, exemplificando, o procedimento de Storey declara 2430 pares significativos com proporção esperada de encontrar um falso positivo nunca superior a 0.01.

Para aplicar o procedimento de Storey, estimou-se a proporção das hipóteses verdadeiras no total das 3904 consideradas, ($\widehat{\pi}_0$), valor obtido a partir de *software* desenvolvido pelo autor. Para a espécie *Saccharomyces cerevisiae* obteve-se o valor 0.22841 e para a *Candida albicans* o valor 0.1632. Isto implica seja grande a proporção estimada de hipóteses significativas.

Relativamente ao procedimento de Storey, o referido *software* permite, de modo gráfico relacionar os valores de $p-value$, $q-value$, número de hipóteses significativas entre as 3904 e a pFDR, conforme Figuras 4.3 e 4.4. Por exemplo, para a *Saccharomyces cerevisiae*, sem rigor analítico, visualiza-se que para um próximo de 2500 testes significantes se tem um proporção de falsas descobertas de 0.01 que, em 3904 hipóteses, corresponde a cerca de 40 falsos positivos esperados como indica o último gráfico.

Nas Figuras 4.5 e 4.6 apresentam-se os mapas de contexto obtidos a partir de cada um dos procedimentos para a *Saccharomyces cerevisiae*, para a *Candida albicans* os mapas de contexto são os das Figuras 4.7 e 4.8. Para ambas as espécies, há uma mudança nítida entre os mapas de contextos obtidos pelos procedimentos de Storey e Bonferroni, ou seja, para este último o mapa apresenta mais células a preto, o que é esperado pela análise efectuada até aqui. Fica em aberto saber se as regras de associação obtidas para estes mapas diferem

i	Par codão	STAR	p-value	IND	BF	B-H	ST
1	CAA - CAA	105.216	0	S	S	S	S
2	GCU - GCU	87.115	0	S	S	S	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
849	UUC - UAC	4.71	0.000002	S	S	S	S
850	AUA - AGG	4.7	0.000002	S	N	S	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1105	AUU - UCU	2.99	0.002789	S	N	S	S
1106	CAA - CGG	2.981	0.002873	S	N	N	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1178	AGA - AAA	2.582	0.009822	S	N	N	S
1179	GCG - ACA	2.573	0.010082	N	N	N	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1220	UGC - UCC	2.344	0.019078	N	N	N	S
1221	CUG - GAG	2.338	0.019387	N	N	N	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2288	CGA - AAC	-2.096	0.036082	N	N	N	N
2289	GCA - CGC	-2.099	0.035816	N	N	N	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2411	GGG - CGG	-2.574	0.010053	N	N	N	S
2412	AUC - CGG	-2.581	0.009851	S	N	N	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2461	CAA - AGG	-2.73	0.006333	S	N	N	S
2462	CUC - CCA	-2.744	0.006069	S	N	S	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2874	GGC - CUA	-4.687	0.000002	S	N	S	S
2875	AUU - UAC	-4.708	0.000002	S	S	S	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3904	UUU - CCA	-32.691	0	S	S	S	S

Tabela 4.3: Pares de codões significativos para a espécie *Candida Albicans*, segundo os quatro procedimentos: Individual(IND), Bonferroni (BF), Bejamini-Hochberg (B-H) e Storey (ST)) para controlo da taxa de erro correspondente ($p-value$, FWER, FDR e pFDR) com $\alpha = 0.01$. S e N denotam pares significantes e pares não significantes respectivamente.

Procedimento	#pares preferidos	#pares sig. preteridos
Teste Individual	1178	1493
Bonferroni	849	1030
Bejamini-Hochberg	1105	1443
Storey	1220	1616

Tabela 4.4: Resumo dos resultados dos testes simultâneos - da espécie *Candida Albicans*.

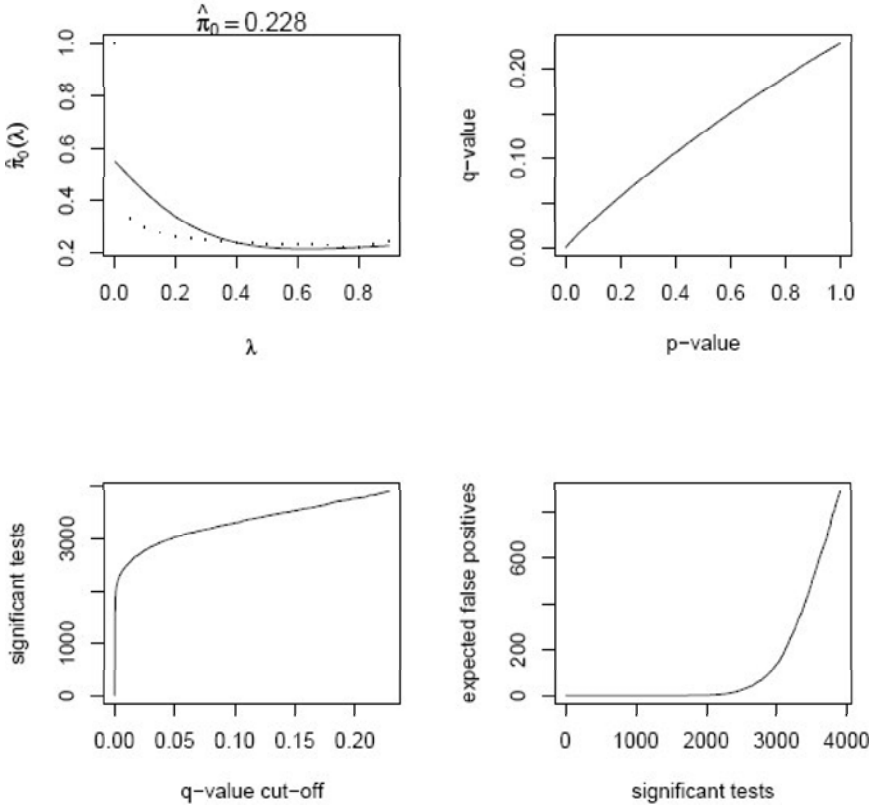


Figura 4.3: Resultados do conjunto de dados de pares de codões da espécie *Saccharomyces cerevisiae*. (a) λ *vs* $\widehat{\pi}_0(\lambda)$. (b) $q - value$ *vs* o respectivo valor $p - values$. (c) número de testes significantes *vs* $q - value$ cut-off. (d) número esperado de falsos positivos *vs* número testes significantes.

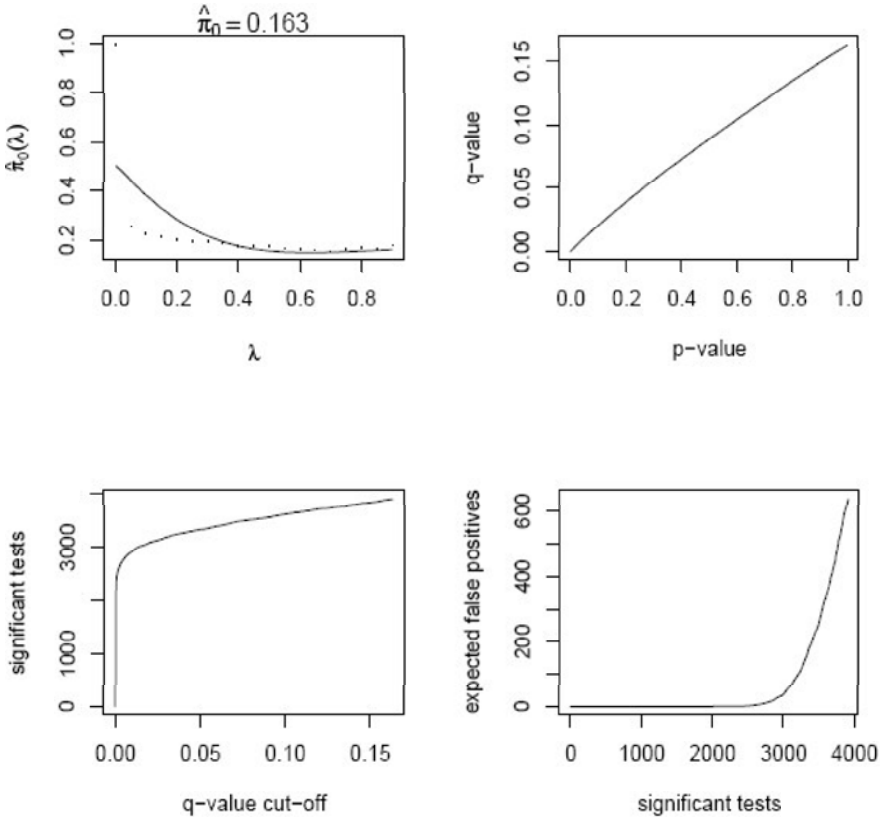


Figura 4.4: Resultados do conjunto de dados de pares de codões da espécie *Candida albicans*. (a) λ *vs* $\widehat{\pi}_0(\lambda)$. (b) $q-value$ *vs* o respectivo valor $p-value$. (c) número de testes significantes *vs* $q-value$ cut-off. (d) número esperado de falsos positivos *vs* número testes significantes.

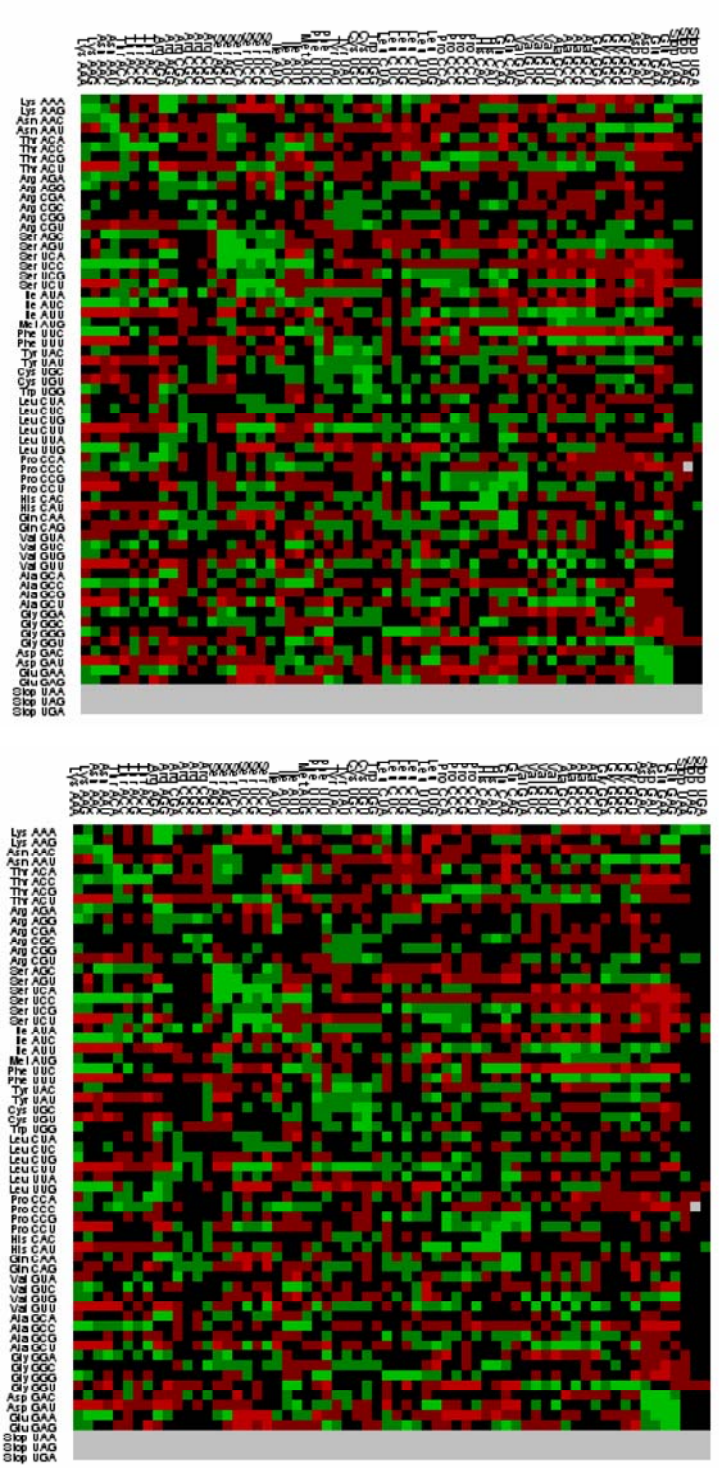


Figura 4.5: Mapas de Contextos $3'$ - Imagem correspondente à matriz dos resíduos da espécie *Saccharomyces cerevisiae* c/ procedimentos de Storey e Bejamini-Hochberg, respectivamente.

perdendo-se informação relevante com a utilização do mapa de Bonferroni.

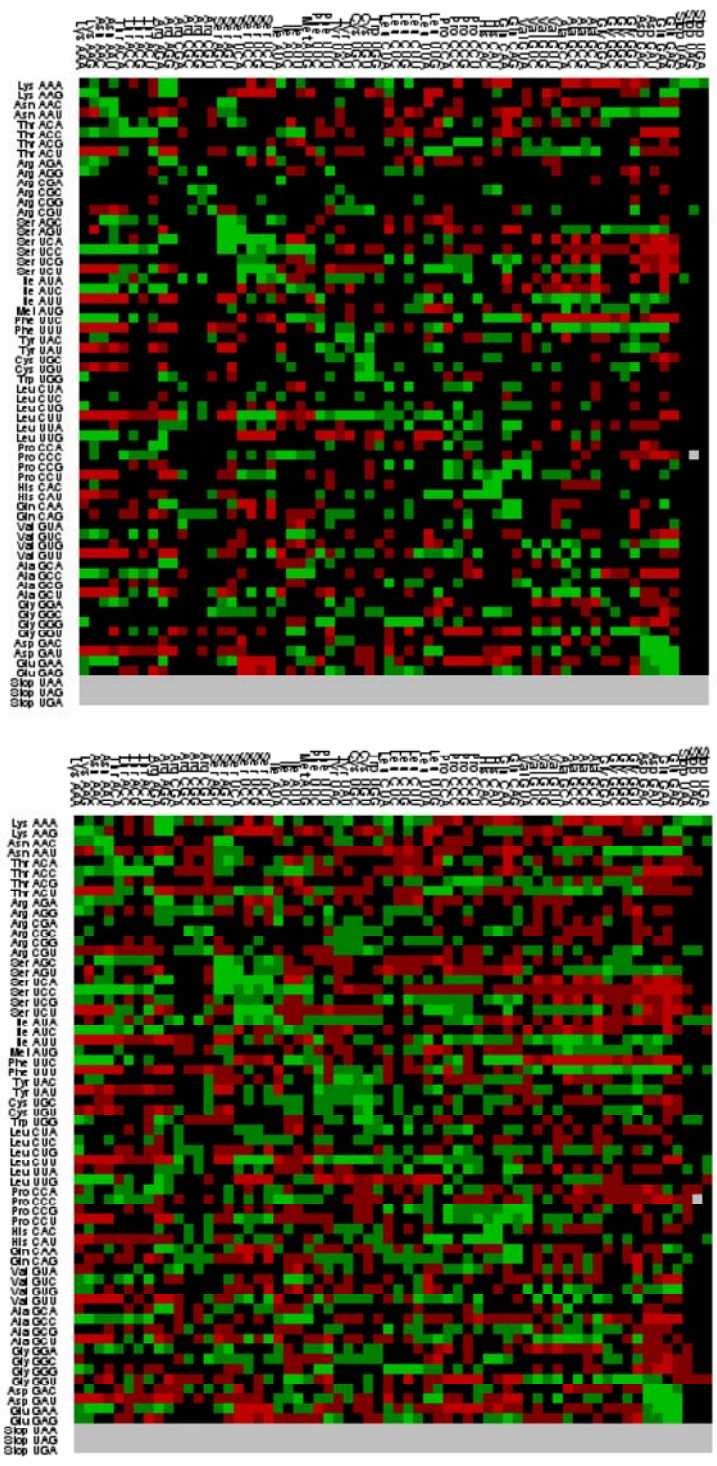


Figura 4.6: Mapas de Contextos 3′ - Imagem correspondente à matriz dos resíduos da espécie *Saccharomyces cerevisiae* c/ procedimentos de Bonferroni e Individual, respectivamente.

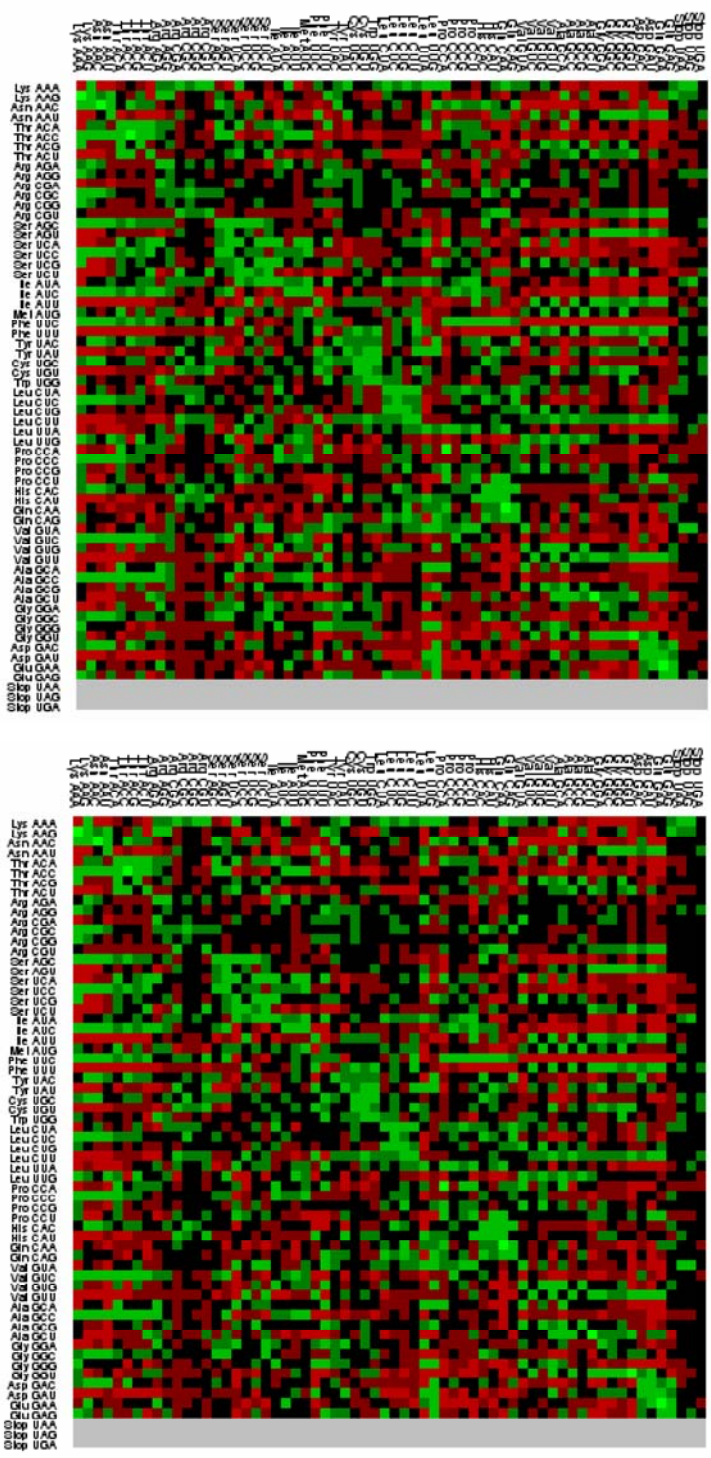


Figura 4.7: Mapas de Contextos 3′ - Imagem correspondente à matriz dos resíduos da espécie *Candida albicans* c/ procedimentos de Storey e Bejamini-Hochberg, respectivamente.

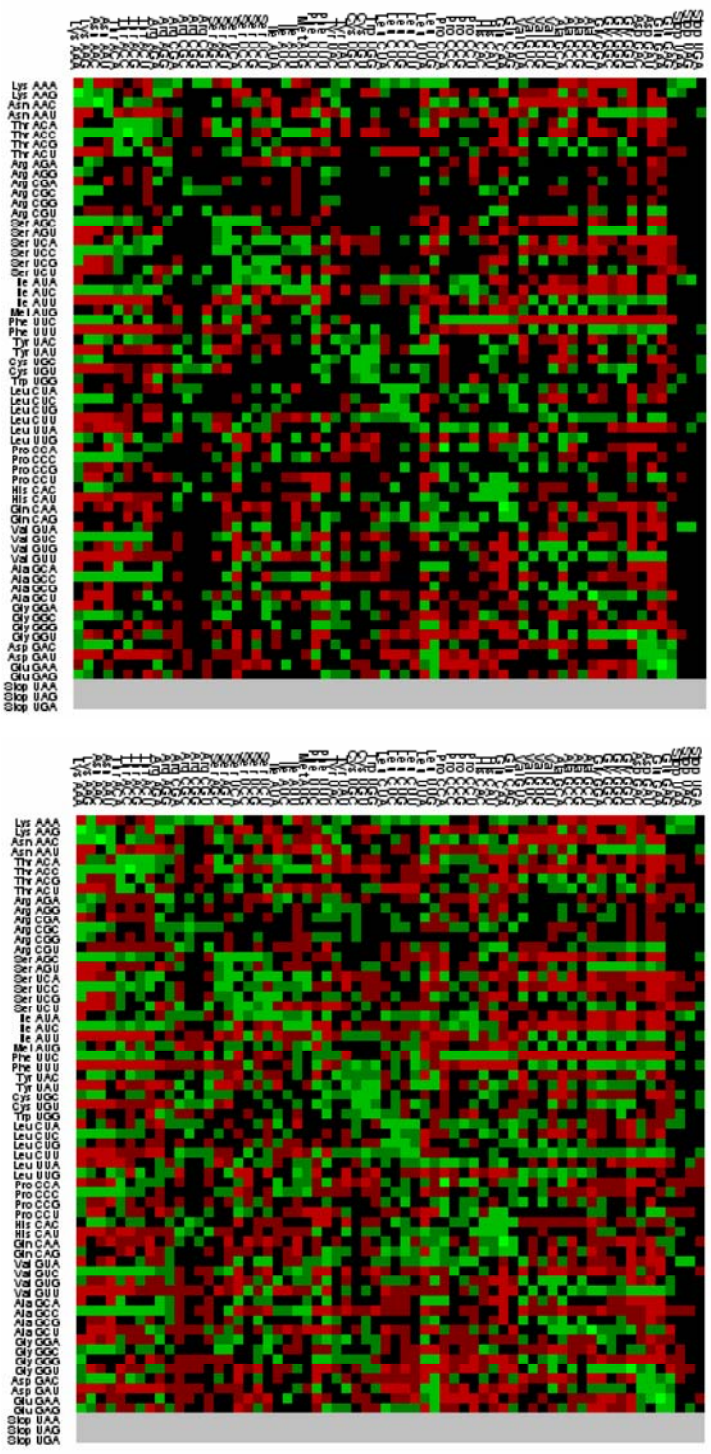


Figura 4.8: Mapas de Contextos $3'$ - Imagem correspondente à matriz dos resíduos da espécie *Candida albicans* c/ procedimentos de Bonferroni e Individual, respectivamente.

4.3 Estimação da proporção de elementos "problemáticos" na diagonal por espécie

Independentemente dos procedimentos atrás considerados para testes simultâneos, todos os mapas salientam uma colorido verde na diagonal das espécies estudadas. Tal observação levanta a questão se os pares de codões iguais definem alguma relação no comportamento no sequenciamento das zonas codificantes do DNA. Propõe-se então uma análise da diagonal principal dos mapas de contextos de pares de codões justapostos para um grupo mais vasto de espécies. Os organismos existentes na Terra dividem-se em três grandes domínios: *Eukarya*, *Bacteria* e *Archaea*. O presente estudo incide sobre um grupo de 119 espécies, as quais se encontram listadas no Apêndice B, sendo 18 *Archaea*, 20 *Eukarya* e 81 *Bacteria*.
Ao longo deste estudo, admite-se que os 61 pares de codões iguais são independentes entre si no sentido em que, o facto de dois codões iguais consecutivos estarem significativamente associados em nada afecta, em termos probabilísticos, a existência ou não de significância de associação entre quaisquer outros pares de codões iguais e justapostos.

IC para a proporção de elementos coloridos na diagonal por espécie

Analisando o número de elementos coloridos na diagonal principal, ie, pares de codões iguais, nos mapas de contextos de pares de codões das 119 espécies. Definindo 119 variáveis aleatórias, uma para cada espécie em estudo, dadas por:

$$Y_j = \text{ número de elementos coloridos na diagonal na espécie } j\,, \qquad j = 1, 2, \cdots 119.$$

tem-se

$$Y_j \sim B(n = 61, p_j),$$

onde p_j é a probabilidade de um elemento da diagonal ser colorido.
Utilizando as metodologias indicadas no Capítulo 2, dado tratarem-se de amostras de grandes dimensões, obtiveram-se as estimativas de máxima verosimilhança de p_j, denotadas por $\hat{p}_j$, $j = 1, 2, \cdots, 119$ e estimativas intervalares denotadas por I_N, I_S e I_{BS}. Os resultados obtidos, usando o mapa de contexto determinado pelo Anaconda, e para todas as 119 espécies, encontra-se em D:\Análise estatística de dados genómicos\Alínea 4.3\Abordagem 4.3.1 - 119 espécies signif ANACONDA.
Para além disso, para as 4 espécies consideradas inicialmente, ampliou-se a presente análise considerando os mapas de contextos obtidos tendo em conta os quatro procedimentos estu-

dados na secção anterior, sendo que os estudos feitos as espécies dos domínios *Archaea* e *Bacteria* se remetem para os respectivos Apêndices C e D.

Na Figura 4.9 apresentam-se as diagonais principais coloridas segundo as diferentes discretizações e na Tabela 4.5 os resultados obtidos para a *Saccharomyces cerevisiae*. De igual modo, para a espécie *Candida Albicans* se mostram a Figura 4.10 e a Tabela 4.6.

Para ambas as espécies o procedimento de Bonferroni não destaca diferenças da estimação no número de elementos coloridos na diagonal, ao passo que os outros métodos em particular o de Storey evidencia codões mais coloridos na diagonal. Esta conclusão vem ao encontro da necessidade de avaliar até que ponto o procedimento de Bonferroni conduz a uma perca significativa de informação.

Analisando as estimativas intervalares obtêm-se valores superiores para os extremos dos intervalos para a *Candida albicans* o que indica que esta espécie apresenta maior número de elemento coloridos na diagonal.

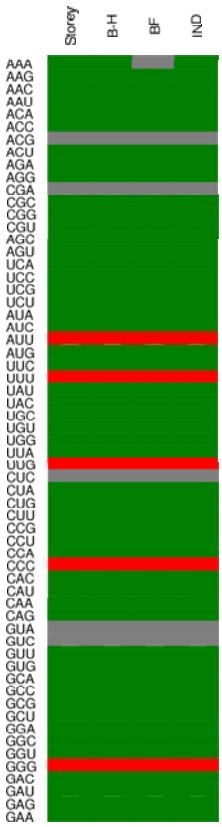


Figura 4.9: Diagonal principal da espécie *Saccharomyces cerevisiae* segundo os procedimentos de ST, B-H, BF e IND, respectivamente.

	Storey	B-H IND	BF
Y	56		55
$\widehat{p_j}$	0.918		0.902
I_N	0.849		0.827
	0.987		0.976
I_S	0.822		0.802
	0.964		0.954
I_{BS}	0.812		0.791
	0.969		0.959

Tabela 4.5: IC a 95% para p_j a probabilidade de um elemento da diagonal ser colorido para a espécie *Saccharomyces cerevisiae*.

	Storey	B-H IND	BF
Y	58	57	55
$\widehat{p_j}$	0.951	0.934	0.902
I_N	0.897	0.872	0.827
	1	0.997	0.976
I_S	0.865	0.843	0.802
	0.983	0.974	0.954
I_{BS}	0.854	0.833	0.791
	0.987	0.979	0.959

Tabela 4.6: IC a 95% para p_j a probabilidade de um elemento da diagonal ser colorido para a espécie *Candida albicans*.

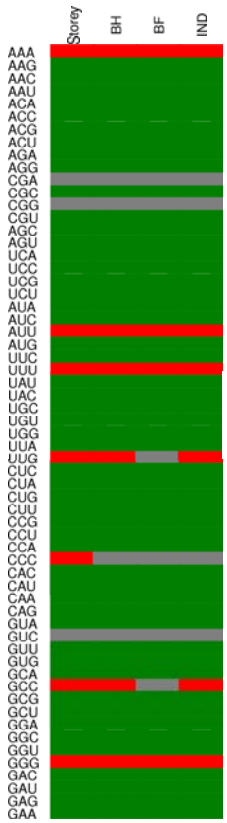


Figura 4.10: Diagonal principal da espécie *Candida albicans* segundo os procedimentos de ST, B-H, BF e IND, respectivamente.

IC a proporção de elementos $r/g/b$ na diagonal por espécie

A abordagem anterior em nada informa se há maior incidência sobre se os pares de codões são preferidos ou preteridos, pelo que se apresenta um estudo mais individualizado.

Considerou-se o número de pares de codões vermelhos (*red*), verdes (*green*) e pretos (*black*) da diagonal.

Analisou-se o número de elementos vermelhos, verdes e pretos na diagonal principal nos mapas de contextos de pares de codões das 119 espécies. Definiram-se 119 vectores tridimensionais de v.a.'s, um para cada espécie em estudo,

$$\mathbf{Y}_j = (Y_r, Y_g, Y_b), \quad j = 1, 2, \cdots 119,$$

onde cada componente representa o número de elementos *red*, *green*, *black* na diagonal na espécie j. Tem-se

$$\mathbf{Y}_j \sim M(n = 61, \mathbf{p}_j = (p_r, p_g, p_b)),$$

onde $\mathbf{p}_j$ é o vector probabilidade de um elemento da diagonal ser $r/g/b$.

Utilizando as metodologias indicadas no Capítulo 2, obtiveram-se as estimativas de máxima verosimilhança de $\mathbf{p}_j = (p_r, p_g, p_b)$, denotadas por $\hat{\mathbf{p}}_j = (\hat{p}_r, \hat{p}_g, \hat{p}_b)$, $j = 1, 2, \cdots, 119$ e estimativas intervalares denominadas por IC Gold, IC Q-Hurst e IC Goodman.

Mostram-se os resultados obtidos para a significância determinada pelo Anaconda para todas as espécies em D:\Análise estatística de dados genómicos \Alínea 4.3 \Abordagem 4.3.2 - 119 espécies signif ANACONDA.

Nas Tabelas 4.7 e 4.8 apresentam-se os resultados obtidos para as espécies *Saccharomyces cerevisiae* e *Candida Albicans*.

Deste estudo mais individualizado destaca-se que, os procedimentos que coloriam a diagonal da mesma forma acabam por colorir os elementos de igual forma, ou seja, o mesmo número de vermelhos, verdes e pretos. Além disso todas as estimativas mostram que a proporção de verdes é muito elevada.

4.4 Estimação da proporção de espécies por pares de codões iguais

IC para a proporção de espécies por pares de codões iguais coloridos

Pretende-se averiguar o comportamento dos pares de codões iguais nos 3 domínios com base nas 119 espécies. A partir deste ponto consideram-se somente os dados resultantes dos mapas de contextos construídos pelo Anaconda.

Analisou-se o número de espécies que apresentam coloração por cada par de codão da diagonal principal.

Para cada domínio definem-se 61 v.a.'s, uma para cada par de codão da diagonal, dadas por:

$$Y_j = \text{número de espécies que apresentam coloração no par de codão igual } j\,, \quad j = 1, 2, \cdots 61.$$

Assume-se que

$$Y_j \sim B(n,\ p_j),$$

onde p_j é a probabilidade de uma espécie apresentar coloração para o par de codões iguais j, sendo $n = 18$ para o domínio dos *Archaea*, $n = 20$ para o domínio dos *Eukarya* e $n = 81$ para o domínio dos *Bacteria*.

Utilizando metodologias indicadas no Capítulo 2, obteve-se a estimativa de máxima verosimilhança de p_j, denotada por $\hat{p}_j$ e estimativas intervalares. As amostras relativas aos domínios

	Storey	
	B-H IND	BF
Y_r	5	5
Y_g	51	50
Y_b	5	6
$\widehat{p}_r$	0.082	0.082
$\widehat{p}_g$	0.836	0.820
$\widehat{p}_b$	0.082	0.098
IC Gold		
L_r	-0.004	-0.004
U_r	0.168	0.168
L_g	0.720	0.699
U_g	0.952	0.940
L_b	-0.004	0.005
U_b	0.168	0.192
IC Q-Hurst		
L_r	0.029	0.029
U_r	0.210	0.210
L_g	0.691	0.673
U_g	0.921	0.910
L_b	0.029	0.038
U_b	0.210	0.230
IC Goodman		
L_r	0.03	0.03
U_r	0.206	0.206
L_g	0.695	0.676
U_g	0.919	0.908
L_b	0.03	0.039
U_b	0.210	0.227

Tabela 4.7: IC a 95% para $\mathbf{p}_j = (p_r, p_g, p_b)$ a probabilidade de um elemento da diagonal ser $r/g/b$ para a espécie *Saccharomyces cerevisiae.*

	Storey	B-H IND	BF
Y_r	7	6	4
Y_g	51	51	51
Y_b	3	4	6
$\widehat{p}_r$	0.115	0.098	0.066
$\widehat{p}_g$	0.836	0.836	0.836
$\widehat{p}_b$	0.049	0.066	0.098
IC Gold			
L_r	0.015	0.005	-0.012
U_r	0.215	0.192	0.143
L_g	0.72	0.72	0.72
U_g	0.952	0.952	0.952
L_b	-0.019	-0.012	0.005
U_b	0.117	0.143	0.192
IC Q-Hurst			
L_r	0.048	0.038	0.021
U_r	0.251	0.23	0.188
L_g	0.691	0.691	0.691
U_g	0.921	0.921	0.921
L_b	0.013	0.021	0.038
U_b	0.166	0.188	0.230
IC Goodman			
L_r	0.049	0.039	0.021
U_r	0.247	0.227	0.184
L_g	0.695	0.695	0.695
U_g	0.919	0.919	0.919
L_b	0.014	0.021	0.039
U_b	0.162	0.184	0.227

Tabela 4.8: IC a 95% para $\mathbf{p}_j = (p_r, p_g, p_b)$ a probabilidade de um elemento da diagonal ser $r/g/b$ para a espécie *Candida albicans*.

Archaea e *Eukarya* são consideradas de pequena dimensão ($n \leq 30$) logo, determinaram-se os IC de Clopper/Pearson e Blyth/Still, e para o caso do domínio *Bacteria* calcularam-se os IC: I_N, I_S e I_{BS}.

Mostram-se os resultados obtidos para todas as espécies em D:\Análise estatística de dados genómicos \Alínea 4.4 \Abordagem 4.4.1.

Nas Tabelas 4.9, 4.10 e 4.11 apresentam-se os resultados mais importantes, onde os pares de codões estão ordenados, por ordem decrescente, pelo número de espécies para as quais estes são coloridos, obtidos para os três domínios.

Na Figura 4.11 é apresentada uma visualização dos dados tomando em conta os IC Clopper e Pearson para os domínios dos *Eukarya* e *Archaea*, e IC I_{BS} para o domínio dos *Bacteria*.

É no domínio dos *Eukarya* onde se visualiza maior número de pares de codões iguais com maiores estimativas intervalares excepção à regra para os codões ACG e GCA. No domínio dos *Bacteria* é onde os intervalos apresentam menor amplitude, ao contrário do domínio dos *Archaea* sendo este último onde se observam menores limites inferiores para estimativas da proporção de codões iguais coloridos, provavelmente indiciando uma não tipificação da existência da diagonal muito colorida ao contrário do domínio dos *Eukarya*.

IC para a proporção de espécies por pares de codões iguais $r/g/b$

Apresenta-se, agora, um estudo mais individualizado para se conseguir averiguar a incidência de pares de codões iguais $r/g/b$ em cada domínio.

Por cada domínio, definem-se 61 vectores tridimensionais de v.a./s, um para cada par de codão da diagonal, dados por:

$$\mathbf{Y}_j = (Y_r, Y_g, Y_b) \qquad j = 1, 2, \cdots 61,$$

onde Y_r,Y_g e Y_b representam o número de espécies c/ coloração *red*, *green*, *black*, respectivamente, no par de codão igual j.

Considera-se

$$\mathbf{Y}_j \sim M(n, \mathbf{p}_j = (p_r, p_g, p_b)),$$

onde $\mathbf{p}_j$ é o vector probabilidade de uma espécie apresentar coloração $r/g/b$ por cada par de codão j da diagonal, sendo $n = 18$ para o domínio dos *Archaea*, $n = 20$ para o domínio dos *Eukarya* e $n = 81$ para o domínio dos *Bacteria*.

Apresentam-se os resultados obtidos para todas as espécies em D:\Análise estatística de dados genómicos \Alínea 4.4 \Abordagem 4.4.2.

	Y_j	$\widehat{p}_j$	IC Clopper e Pearson		IC Blyth/Still	
AAG, AAC, ACC, AGA AGC, UCA, UCC, UCG UCU, AUA, AUC, UGC CCU, CCA, CAU, CAA CAG, GUU, GCA, GCU GGU, GAG, GAA	20	1	0.831	1	0.84	1
AGG, CGC, UUC, UAC UUA, CUU, CCG, GUG GCG, GGA, GAU	19	0.95	0.751	0.999	0.76	1
CGU, AGU, UUU, CUG CAC, GCC, GGC, GAC	18	0.90	0.683	0.988	0.68	0.98
AAA, AAU, ACU, CGG UAU, UGU, CUC, GUC GGG	17	0.85	0.621	0.968	0.63	0.96
UGG, CCC	16	0.80	0.563	0.943	0.58	0.93
ACA, AUU	15	0.75	0.509	0.913	0.53	0.90
AUG, CUA	14	0.7	0.457	0.881	0.47	0.86
UUG, GUA	13	0.65	0.408	0.846	0.42	0.84
ACG	11	0.55	0.315	0.769	0.32	0.76
CGA	10	0.5	0.272	0.728	0.29	0.71

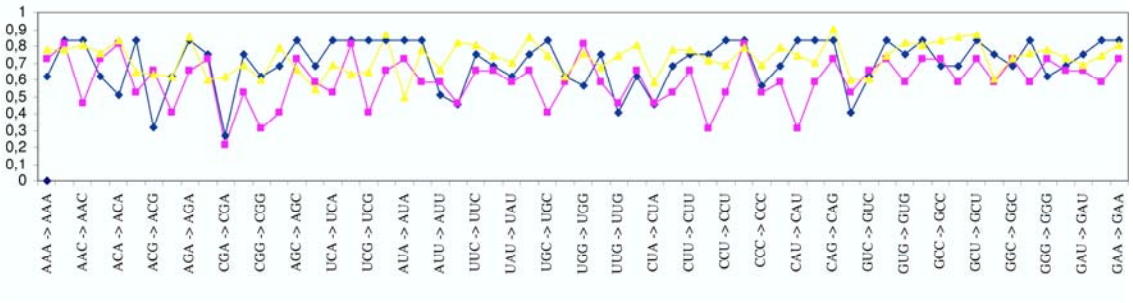
Tabela 4.9: IC a 95% para p_j a probabilidade de uma espécie, do domínio *Eukarya*, ser colorida para um par de codões j.

	Y	$\widehat{p}_j$	IC Clopper e Pearson		IC Blyth/Still	
AAG, ACA UCC, UGG CCA	18	1	0.816	1	0.82	1
AAA, AAU, AGG, AGC AUA, CAG, GUU, GCA GCC, GCU, GGC, GGG GAA	17	0.944	0.727	0.999	0.73	1
ACG, AGA, UCU, UUC UUU, UAC, CUC, CUU GUC, GAC, GAU	16	0.888	0.652	0.986	0.67	0.98
AGU, AUC, AUU, UAU UGU, UUA, CAC, CAA GUG, GCG, GGA, GGU GAG	15	0.833	0.586	0.964	0.59	0.95
ACC, CGC, UCA, CUG CCU, CCC, GUA	14	0.777	0.524	0.936	0.53	0.92
AAC, AUG, UUG, CUA	13	0.722	0.465	0.903	0.47	0.88
ACU, CGU, UCG, UGC	12	0.666	0.41	0.867	0.41	0.84
CGG, CCG, CAU	10	0.556	0.308	0.785	0.33	0.76
CGA	8	0.444	0.215	0.692	0.24	0.67

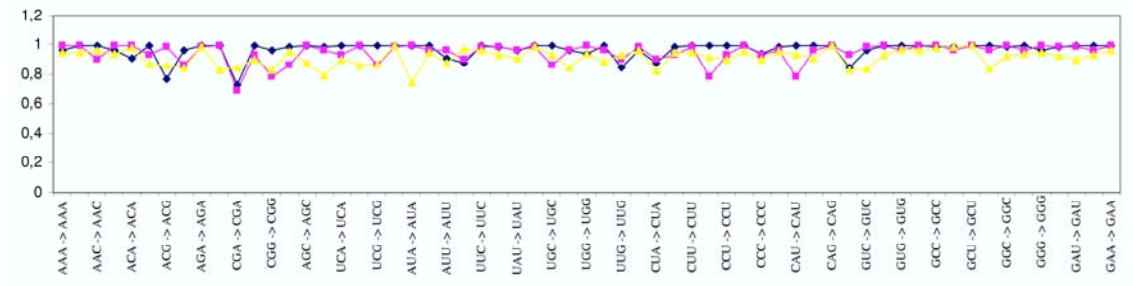
Tabela 4.10: IC a 95% para p_j a probabilidade de uma espécie, do domínio *Archaea*, ser colorida para um par de codões j.

	Y	$\widehat{p}_j$	I_N		I_S		I_{BS}	
CAG	80	0.988	0.96	1	0.919	0.998	0.937	0.999
UCU, GCU	78	0.963	0.916	1	0.882	0.989	0.889	0.993
AGA, UAC, GCG	77	0.951	0.896	1	0.865	0.983	0.87	0.987
ACA, GCC	76	0.938	0.878	0.999	0.848	0.976	0.852	0.981
ACA, GCC	75	0.926	0.861	0.992	0.832	0.97	0.835	0.974
AAC, UUC, CUC, GCA, GAA	74	0.914	0.844	0.984	0.816	0.962	0.818	0.967
CGU, CCA, CAC	73	0.901	0.826	0.976	0.801	0.954	0.802	0.959
AAA, AAG, AUC, CUG, CUU, GGG	72	0.889	0.810	0.968	0.786	0.946	0.787	0.951
AAU, UGG, GGU	71	0.877	0.794	0.959	0.771	0.937	0.771	0.943
UUU, UGC, UUG, CAU, GUU, GAG	70	0.864	0.778	0.950	0.765	0.929	0.756	0.935
GGC, GAC	69	0.852	0.763	0.941	0.742	0.92	0.741	0.926
CCG	68	0.84	0.747	0.932	0.728	0.911	0.727	0.917
UAU, CAA	67	0.827	0.732	0.922	0.714	0.902	0.713	0.908
CAA	67	0.827	0.732	0.922	0.713	0.902	0.713	0.908
CGC, UCA, CCU, CCC, GAU	66	0.815	0.717	0.912	0.7	0.893	0.691	0.899
UUA	65	0.802	0.702	0.902	0.686	0.883	0.677	0.89
AGC, AUU	64	0.790	0.688	0.892	0.672	0.874	0.671	0.88
ACC, UCG	63	0.778	0.673	0.882	0.659	0.864	0.657	0.870
ACG, UCC	62	0.765	0.66	0.872	0.645	0.854	0.643	0.861
ACU, CGA, CGA,UGU	61	0.753	0.645	0.861	0.632	0.844	0.623	0.851
GUC, GGA	60	0.741	0.631	0.851	0.619	0.834	0.61	0.841
AGG, CGG, GUA	59	0.728	0.617	0.84	0.606	0.824	0.597	0.831
CUA	58	0.716	0.603	0.829	0.593	0.814	0.584	0.821
AGU	55	0.679	0.562	0.796	0.554	0.782	0.546	0.79
AUA	51	0.63	0.508	0.751	0.504	0.74	0.496	0.747

Tabela 4.11: IC a 95% para p_j a probabilidade de uma espécie, do domínio *Bacteria*, ser colorida para um par de codões j.



(a)



(b)

Figura 4.11: Limite inferior (a) e limite superior (b) dos IC Clopper e Pearson e IC I_{BS} de acordo com as Tabelas 4.9, 4.10 e 4.11. A azul o domínio dos *Eukarya*, a rosa o dos *Archaea* e a amarelo o domínio dos *Bacteria*.

	UUU	CCC	GUA	AUG	GAG
Y_r	15	13	0	0	0
Y_g	3	4	12	14	20
Y_b	2	3	8	6	0
$\widehat{p}_r$	0.75	0.65	0	0	0
$\widehat{p}_g$	0.15	0.2	0.6	0.7	1
$\widehat{p}_b$	0.1	0.15	0.4	0.3	0
IC Gold					
L_r	0.513	0.389	0	0	0
U_r	0.987	0.911	0	0	0
L_g	-0.045	-0.019	0.332	0.449	1
U_g	0.345	0.419	0.868	0.951	1
L_b	-0.064	-0.045	0.131	0.049	0
U_b	0.264	0.345	0.668	0.550	0
IC Q-Hurst					
L_r	0.477	0.384	0	0	0
U_r	0.908	0.847	0.230	0.231	0.231
L_g	0.041	0.065	0.341	0.429	0.769
U_g	0.420	0.473	0.813	0.879	1
L_b	0.021	0.041	0.187	0.121	0
U_b	0.363	0.420	0.659	0.571	0.231
IC Goodman					
L_r	0.483	0.389	0	0	0
U_r	0.906	0.844	0.222	0.222	0.223
L_g	0.042	0.067	0.345	0.435	0.777
U_g	0.414	0.467	0.810	0.876	1
L_b	0.022	0.042	0.19	0.124	0
U_b	0.356	0.414	0.655	0.566	0.222

Tabela 4.12: IC a 95% para $\mathbf{p}_j = (p_r, p_g, p_b)$ a probabilidade de uma espécie, do domínio *Eukarya*, ser $r/g/b$ para um par de codões iguais j.

Utilizando as metodologias já aplicadas em secções anteriores, obtiveram-se as estimativas de máxima verosimilhança de $\mathbf{p}_j = (p_r, p_g, p_b)$, denotadas por $\hat{\mathbf{p}}_j = (\hat{p}_r, \hat{p}_g, \hat{p}_b)$, $j = 1, 2, \cdots, 61$ e estimativas intervalares denominadas por IC Gold, IC Q-Hurst e IC Goodman.

Nas Tabelas 4.12, 4.13 e 4.14 apresentam-se os resultados extremos relativamente a codões mais *red* e *green*, curiosamente destaca-se o par de codão CCC-CCC mais vermelho nos três domínios

	GGG	CCC	CGA	UGG
Y_r	17	14	0	0
Y_g	0	0	18	18
Y_b	1	4	10	0
$\widehat{p}_r$	0.944	0.778	0	0
$\widehat{p}_g$	0	0	0.444	1
$\widehat{p}_b$	0.056	0.222	0.556	0
IC Gold				
L_r	0.812	0.538	0	0
U_r	1	1	0	0
L_g	0	0	0.158	1
U_g	0	0	0.731	1
L_b	-0.077	-0.018	0.269	0
U_b	0.188	0.462	0.842	0
IC Q-Hurst				
L_r	0.674	0.489	0	0
U_r	0.993	0.927	0.25	0.25
L_g	0	0	0.21	0.75
U_g	0.25	0.25	0.707	1
L_b	0.007	0.073	0.293	0
U_b	0.326	0.511	0.79	0
IC Goodman				
L_r	0.682	0.496	0	0
U_r	0.993	0.926	0.242	0.242
L_g	0	0	0.198	0.67
U_g	0.242	0.242	0.702	1
L_b	0.007	0.074	0.298	0
U_b	0.318	0.504	0.787	0.242

Tabela 4.13: IC a 95% para $\mathbf{p}_j = (p_r, p_g, p_b)$ a probabilidade de uma espécie, do domínio *Archaea*, ser $r/g/b$ para um par de codões iguais j.

	CUC	CCC	AUA	CGG	AGA
Y_r	60	59	20	6	0
Y_g	14	7	31	53	77
Y_b	7	15	30	22	4
p_r	0.741	0.729	0.247	0.074	0
p_g	0.173	0.086	0.383	0.654	0.951
p_b	0.087	0.185	0.370	0.272	0.049
IC Gold					
L_r	0.622	0.607	0.129	0.003	0
U_r	0.86	0.849	0.364	0.145	0
L_g	0.07	0.01	0.251	0.525	0.892
U_g	0.276	0.163	0.515	0.784	1
L_b	0.01	0.08	0.239	0.151	-0.01
U_b	0.163	0.291	0.502	0.393	0.108
IC Q-Hurst					
L_r	0.608	0.595	0.15	0.029	0
U_r	0.840	0.830	0.379	0.178	0.069
L_g	0.094	0.036	0.263	0.518	0.855
U_g	0.297	0.194	0.519	0.769	0.984
L_b	0.036	0.103	0.252	0.17	0.016
U_b	0.103	0.311	0.506	0.41	0.145
IC Goodman					
L_r	0.611	0.598	0.152	0.029	0
U_r	0.839	0.827	0.376	0.175	0.066
L_g	0.095	0.037	0.265	0.521	0.858
U_g	0.294	0.191	0.516	0.767	0.984
L_b	0.037	0.104	0.255	0.171	0.016
U_b	0.191	0.308	0.503	0.402	0.142

Tabela 4.14: IC a 95% para $\mathbf{p}_j = (p_r, p_g, p_b)$ a probabilidade de uma espécie, do domínio *Bacteria*, ser $r/g/b$ para um par de codões iguais j.

4.5 Estimação da diferenças de proporções

IC para a diferença de proporções de duas binomiais

Nesta secção pretende-se comparar a diagonal dos mapas de contextos para vários pares de espécies que pertencem à mesma árvore filogenética.

Para o domínio *Eukarya* estudaram-se as espécies *S.cerevisiae vs S.mykatae*, *S.pombe vs S.mykatae*, *S.cerevisiae vs S.pombe*, *S.cerevisiae vs C.albicans* e *H.sapiens vs P.troglodytes.*

No domínio *Archaea* consideraram-se as espécies *T.Kodakaraensis vs T.volcanium*, *T.acidophilum vs T.Kodakaraensis* e *T.volcanium vs T.acidophilum.*

Para o domínio *Bacteria* seleccionaram-se *G.kaustophilus vs O.iheyensis*, *G.kaustophilus vs B.cereus* e *O.iheyensis vs B.cereus.*

Em cada domínio, e para cada par de espécies consideradas, pretendeu-se comparar a diferença de proporções de elementos coloridos existentes na diagonal dos mapas de contextos entre cada uma.

Para tal tomaram-se as v.a.'s Binomiais definidas na secção 4.3 e encontraram-se estimativas intervalares com grau de confiança 95%.

Os resultados obtidos encontram-se em D:\Análise estatística de dados genómicos \Alínea 4.5. \Alínea 4.5.1

Utilizaram-se as metodologias indicadas no Capítulo 2, obtiveram-se as estimativas de máxima verosimilhança para a diferença de proporções entre 2 espécies (Espécie 1 *vs* Espécie 2), denotadas nas tabelas seguintes por $\hat{p}_1 - \hat{p}_2$, e estimativas intervalares dadas pelos IC de Wald, Yule, Newcombe e Recentrado.

Nas Tabelas 4.15, 4.16 e 4.17 apresentam-se os resultados obtidos para três domínios.

Com base no IC pode-se testar a hipótese de igualdade de proporções entre 2 espécies a partir do facto do número zero pertencer ou não ao IC. Assim sendo, da análise das referidas tabelas, pode inferir-se que, em termos do comportamento de coloração dos pares de codões iguais, não existem razões para rejeitar aquelas hipóteses nulas para os casos H.sapiens *vs* P.troglodytes, S.cerevisiae *vs* S.pombe e S.cerevisiae *vs* C.albicans.

Nos domínios dos *Archaea* e *Bacteria*, pelos resultados obtidos não existem razões para rejeitar o mesmo tipo de hipótese nula para os casos estudados.

O caso mais curioso será o do par S.cerevisiae *vs* C.albicans dado que se consideram que estas espécies degeneraram na sua árvore filogenética.

	H.sapiens vs P.troglodytes	*S.cerevisiae vs S.mykatae*	*S.cerevisiae vs S.pombe*	*S.pombe vs S.mykatae*	*S.cerevisiae vs C.albicans*
$\widehat{p}_1 - \widehat{p}_2$	0.016	0.180	-0.016	0.197	0
IC Wald	-0.093	0.453	-0.118	0.065	-0.106
	0.126	0.316	0.085	0.033	0.106
IC Yule	-0.093	0.042	-0.118	0.060	-0.106
	0.126	0.319	0.085	0.333	0.106
IC Newcombe	-0.099	0.041	-0.127	0.060	-0.113
	0.133	0.314	0.093	0.328	0.113
IC Recentrado	-0.092	0.042	-0.116	0.061	-0.104
	0.124	0.308	0.084	0.321	0.104

Tabela 4.15: IC a 95% para a diferença de proporções de elementos coloridos na diagonal dos mapas de contextos para estas espécies do domínio *Eukarya*.

	T.Kodakaraensis vs T.volcanium	*T.volcanium vs T.acidophilum*	*T.acidophilum vs T.Kodakaraensis*
$\widehat{p}_1 - \widehat{p}_2$	0.082	0	0.082
IC Wald	-0.04	-0.014	-0.04
	0.204	0.137	0.204
IC Yule	-0.041	-0.014	-0.041
	0.205	0.137	0.205
IC Newcombe	-0.044	-0.014	-0.044
	0.208	0.137	0.208
IC Recentrado	-0.041	-0.014	-0.041
	0.2	0.137	0.2

Tabela 4.16: IC a 95% para a diferença de proporções de elementos coloridos na diagonal dos mapas de contextos para estas espécies do domínio *Archaea*.

IC para a diferença de proporções de duas multinomiais

Considerando os mesmos moldes da secção anterior, neste ponto pretendeu-se um estudo mais individualizado para se conseguir averiguar a diferença de proporções pelas componentes $r/g/b$.

Os resultados obtidos usando estimativas intervalares dadas por Gold e Goodman, encontram-se em D:\Análise estatística de dados genómicos\Alínea 4.5.\Alínea 4.5.2. Nas Tabelas 4.18, 4.19 e 4.20 apresenta-se uma súmula para os três domínios.

No domínio dos *Eukarya*, para o par S.pombe *vs* S.mykatae pode-se concluir que a coloração black é significativamente diferente. Nos outros domínios tudo leva a concluir pela não rejeição da hipótese de igualdade de proporção qualquer que seja a coloração.

4.6 Estimação da proporção de pares de codões iguais

Até este ponto o estudo incidiu sobre o número de elementos coloridos na diagonal dos mapas de contextos. A coloração depende do nível do valor crítico considerado para a estatística STAR. A fim de analisar a diagonal sem ter em conta os valores da referida estatística procedeu-se a uma análise da frequência absoluta dos pares de codões iguais.

Nesta análise restringiram-se o número de espécies tendo sido consideradas as espécies: *G.kaustophilus*, *O.iheyensis*, *B.cereus* (do domínio dos *Bacteria*), *T.Kodakaraensis*, *T.volcanium*, *T.acidophilum* (do domínio dos *Archaea*), *H.sapiens*, *S.cerevisiae*, *S.pombe*, *P.troglodytes*, *S.mykatae* e *C.albicans* (do domínio dos *Eukarya*).

	G.kaustophilus vs O.iheyensis	*O.iheyensis vs B.cereus*	*G.kaustophilus vs B.cereus*
$\widehat{p}_1 - \widehat{p}_2$	0.033	-0.049	-0.082
IC Wald	-0.093	-0.019	-0.215
	0.159	0.089	0.051
IC Yule	-0.093	-0.188	-0.216
	0.159	0.09	0.052
IC Newcombe	-0.096	-0.188	-0.216
	0.162	0.091	0.054
IC Recentrado	-0.092	-0.184	-0.211
	0.156	0.089	0.052

Tabela 4.17: IC a 95% para a diferença de proporções de elementos coloridos na diagonal dos mapas de contextos para estas espécies do domínio *Bacteria*.

	H.sapiens *vs* P.troglodytes	S.cerevisiae *vs* S.mykatae	S.cerevisiae *vs* S.pombe	S.pombe *vs* S.mykatae	S.cerevisiae *vs* C.albicans
$\widehat{p}_{1_r} - \widehat{p}_{2_r}$	0	0.066	-0.033	0.098	0.016
$\widehat{p}_{1_g} - \widehat{p}_{2_g}$	0.016	0.115	0.016	0.098	-0.016
$\widehat{p}_{1_b} - \widehat{p}_{2_b}$	-0.016	-0.18	0.016	-0.197	0
IC Gold					
L_r	-0.188	-0.031	-0.171	-0.011	-0.103
U_r	0.188	0.162	0.106	0.208	0.137
L_g	-0.358	-0.271	-0.383	-0.285	-0.419
U_g	0.391	0.5	0.416	0.482	0.387
L_b	-0.161	-0.364	-0.117	-0.375	-0.139
U_b	0.128	0.004	0.149	-0.019	0.139
IC Goodman					
L_r	-0.184	-0.028	-0.168	-0.008	-0.101
U_r	0.184	0.16	0.103	0.205	0.134
L_g	-0.35	-0.26	-0.374	-0.277	-0.411
U_g	0.382	0.492	0.407	0.473	0.378
L_b	-0.158	-0.36	-0.114	-0.371	-0.136
U_b	0.125	0	0.146	-0.023	0.136

Tabela 4.18: IC a 95% para $\mathbf{p}_{1_j} - \mathbf{p}_{2_j}, \quad j = r/g/b$ a probabilidade da diferença de um elemento da diagonal ser $r/g/b$ para um par de espécies do domínio *Eukarya*.

	T.Kodakaraensis *vs* T.volcanium	T.volcanium *vs* T.acidophilum	T.acidophilum *vs* T.Kodakaraensis
$\widehat{p}_{1_r} - \widehat{p}_{2_r}$	-0.115	0.066	-0.18
$\widehat{p}_{1_g} - \widehat{p}_{2_g}$	0.198	-0.066	0.262
$\widehat{p}_{1_b} - \widehat{p}_{2_b}$	-0.082	0	-0.082
IC Gold			
L_r	-0.363	-0.217	-0.437
U_r	0.133	0.349	0.0767
L_g	-0.123	-0.349	-0.043
U_g	0.517	0.217	0.567
L_b	-0.245	-0.189	-0.245
U_b	0.0815	0.188	0.081
IC Goodman			
L_r	-0.357	-0.211	-0.432
U_r	0.128	0.342	0.071
L_g	-0.116	-0.342	-0.036
U_g	0.51	0.211	0.561
L_b	-0.242	-0.184	-0.242
U_b	0.078	0.184	0.078

Tabela 4.19: IC a 95% para $\mathbf{p}_{1_j} - \mathbf{p}_{2_j}, \quad j = r/g/b$ a probabilidade da diferença de um elemento da diagonal ser $r/g/b$ para um par de espécies do domínio *Archea*.

	G.kaustophilus *vs* O.iheyensis	O.iheyensis *vs* B.cereus	G.kaustophilus *vs* B.cereus
$\widehat{p}_{1_r} - \widehat{p}_{2_r}$	-0.066	0	0.066
$\widehat{p}_{1_g} - \widehat{p}_{2_g}$	0.098	-0.049	-0.148
$\widehat{p}_{1_b} - \widehat{p}_{2_b}$	-0.033	0.049	0.082
IC Gold			
L_r	-0.277	-0.227	-0.146
U_r	0.146	0.227	0.277
L_g	-0.250	-0.377	-0.487
U_g	0.447	0.279	0.192
L_b	-0.203	-0.143	-0.100
U_b	0.137	0.241	0.264
IC Goodman			
L_r	-0.272	-0.222	-0.141
U_r	0.141	0.222	0.272
L_g	-0.242	-0.370	-0.48
U_g	0.439	0.272	0.185
L_b	-0.199	-0.138	-0.096
U_b	0.133	0.237	0.26

Tabela 4.20: IC a 95% para $p_{1_j} - p_{2_j}$, $\quad j = r/g/b$ a probabilidade da diferença de um elemento da diagonal ser $r/g/b$ para um par de espécies do domínio *Bacteria*.

Assim, definiram-se 12 v.a.′s, uma por cada espécie em estudo, dadas por:

$$Z_i = \text{número de pares de codões iguais na espécie } i, \quad i = 1, 2, \cdots, 12.$$

Assume-se que

$$Z_i \sim B(n, p_i),$$

onde p_i é a probabilidade de um par de codões pertencer à diagonal, ie, o par ser composto por dois codões iguais, e o parâmetro n representa o número de pares de codões no genoma da espécie i, uma vez que existem um elevado número de codões por genoma, as metodologias utilizadas para determinar as estimativas intervalares são os IC descritos no Capítulo 2 para amostras de grandes dimensões.

Os resultados obtidos encontram-se em D:\Análise estatística de dados genómicos\Alínea 4.6 e na Figura 4.12.

Da observação dos resultados obtidos constata-se que a estimativa pontual para a proporção p_i é superior ao valor que se esperaria se os pares de codões estivessem uniformemente distribuídos no genoma, ie, $61/3904 = 0.0156$. Por outro lado, analisando as estimativas intervalares verifica-se uma grande diferença entre os IC denotados por I_N, I_S e o IC I_{BS}, as conclusões resultantes contradizem-se. Tendo como base a teoria exposta no Capítulo 2, se se atender às características destas três metodologias dar-se-à maior fiabilidade aos resultados fornecidos por I_{BS}. Assim sendo, e uma vez que o valor 0.0156 não pertence a nenhum destes intervalos, conclui-se pela rejeição da hipótese de uniformidade da distribuição dos pares de codões iguais no genoma.

Uma vez que se tem um elevado número de pares de codões (3904) as estimativas para w pouco diferem das estimativas obtidas para p na realidade se houvesse distribuição uniforme ter-se-ia $w = 61/(3904 - 61) = 0.0159$.

4.7 Análise de codões raros

Nesta secção o objectivo é averiguar se o modelo probabilístico subjacente à frequência de codões raros por gene se ajusta ao modelo de Poisson.

Consideraram-se as seguintes espécies *T.acidophilum* do domínio dos *Archaea*, *B.cereus* do domínio dos *Bacteria*, *S.cerevisiae* e *C.albicans* do domínio dos *Eukarya*.

Começou-se por calcular a média e a variância amostrais para todos os 61 codões (considerados raros ou não) para as 4 espécies. Para cada situação observou-se que a média amostral

		G.Kaustophilus	O.iheyensis	R.cereus	T.kodakaraensis	T.acidophilum	T.volcanium	H.sapiens	S.cerevisiae	S.pombe	P.troglodytes	S.mykatae	C.albicans
N.º de pares de codões na diagonal		22188	24506	32699	14815	7324	8971	363300	86336	58248	126500	52224	143140
N.º de pares de codões fora da diagonal		814990	912911	1247589	544635	367450	399876	12326947	2458407	2082617	4534892	1701021	3229086
Total de pares de codões		837178	937417	1280288	559450	374774	408847	12690247	2544743	2140865	4661392	1753245	3372226
$\widehat{p}$		0.0265	0.0261	0.0255	0.0264	0.0195	0.0219	0.0286	0.0339	0.0272	0.0271	0.0297	0.0424
I_N	-0.014	-0.014	-0.014	-0.014	-0.014	-0.015	-0.015	-0.013	-0.012	-0.014	-0.014	-0.013	-0.008
	0.067	0.066	0.065	0.067	0.054	0.059	0.07	0.079	0.068	0.068	0.072	0.093
I_S	0.006	0.006	0.006	0.006	0.004	0.005	0.007	0.01	0.007	0.007	0.008	0.013
	0.103	0.102	0.101	0.103	0.092	0.096	0.106	0.114	0.104	0.104	0.108	0.126
I_{BS}	0.0262	0.0258	0.0253	0.0261	0.0191	0.0215	0.0285	0.0337	0.027	0.027	0.0295	0.0422
	0.0269	0.0265	0.0258	0.0269	0.02	0.0224	0.0287	0.0342	0.0274	0.0273	0.03	0.0427
$\widehat{w}$		0.0272	0.0268	0.0262	0.0272	0.0199	0.0224	0.0294	0.0351	0.0279	0.0278	0.0307	0.0443
I_N	-0.0138	-0.0138	-0.0138	-0.0138	-0.0138	-0.0147	-0.0147	-0.0128	-0.0118	-0.0138	-0.0138	-0.0128	-0.0079
	0.0718	0.0706	0.0695	0.0718	0.0571	0.0627	0.0752	0.0857	0.0729	0.0729	0.0775	0.01025
I_S	0.006	0.006	0.006	0.006	0.004	0.005	0.007	0.01	0.007	0.007	0.008	0.013
	0.1148	0.1135	0.1123	0.1148	0.1013	0.1061	0.1186	0.1287	0.1161	0.1161	0.1211	0.1442
I_{BS}	0.0269	0.0264	0.0259	0.0267	0.0194	0.0219	0.0293	0.0348	0.0277	0.0277	0.0303	0.0440
	0.0276	0.0272	0.0264	0.0276	0.0204	0.0229	0.0295	0.0354	0.0281	0.0281	0.0309	0.0446

Figura 4.12: IC a 95% para a proporção das frequências absolutas para os pares de codões iguais e para a razão entre sucesso e falha.

Codão	Nº genes em que ocorreu 0 vezes	Nº genes em que ocorreu 49 ou + vezes	Média amostral	Variância amostral
CGA	1999	0	1.518	3.675
CGC	1970	0	1.306	2.577
CGG	2640	0	0.886	1.637
UGC	1211	0	2.378	6.526

Tabela 4.21: Frequência de codões por gene da espécie *S.cervisiae*.

Codão	Nº genes em que ocorreu 0 vezes	Nº genes em que ocorreu 49 ou + vezes	Média amostral	Variância amostral
ACG	3018	0	1.713	5.083
AGG	3664	0	1.353	3.917
CCC	2764	0	1.853	4.839
CCG	3739	0	1.326	3.65
CGC	6246	0	0.413	0.845
CGG	5698	0	0.556	1.341
CUC	3550	0	1.399	4.771
GCC	1556	0	3.873	18.877
GGC	2998	0	1.792	5.353
UGC	4628	0	0.934	2.301

Tabela 4.22: Frequência de codões por gene da espécie *C.albicans*.

não apresenta um valor substancialmente próximo da variância amostral correspondente. As Tabelas 4.21 e 4.22 ilustra a situação para as espécies *S.cervisiae* e *C.albicans*, que possuem 5040 e 8467 genes, respectivamente, e onde se mostram os codões cuja frequência por gene é mais baixa e se comprovam que os valores da média e variância amostrais não são idênticos. Recorrendo ao Anaconda observou-se que se tratam de codões raros, ou seja, codões cuja frequência relativa de ocorrência na sequência do genoma é muito baixa, ie, inferior a 5/1000. Sobre estes aplicaram-se testes gráfico (Método de Hoaglin) e analítico para averiguar o ajustamento ao modelo de Poisson.

Todos os resultados obtidos encontram-se em D:\Análise estatística de dados genómicos \Alínea 4.7.

Ilustra-se aqui o procedimento feito para o codão raro ACG da espécie *C.albicans*. Tomaram-se em conta duas possíveis estimativas para o parâmetro λ.

No primeiro caso realizou-se o teste de ajustamento do Qui-quadrado com $\widehat{\lambda} = \overline{x} = 1.713$. Para $\alpha = 0,05$, $m = 8$ classes e $k = 1$, obteve-se $\chi^2_{obs} = 8572.345 > \chi^2_{0.05,6} = 12.59159$ donde se conclui pela rejeição da hipótese nula de que os dados provêm de uma distribuição de

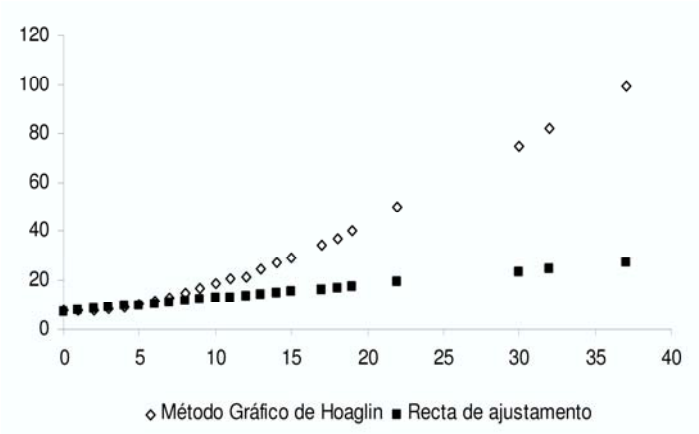


Figura 4.13: Gráfico de Hoaglin para o codão ACG.

Poisson.

O mesmo se passa no caso em que o teste de ajustamento do Qui-quadrado se realizou com $\widehat{\lambda} = s^2 = 5.083$. Assim, para $\alpha = 0,05$, $m = 15$ classes e $k = 1$ obteve-se $\chi^2_{obs} = 184698.7 > \chi^2_{0.05,13} = 22.36203$, volta-se rejeitar a hipótese de que os dados seguem uma distribuição de Poisson.

Com base na Figura 4.13, onde se apresenta o gráfico de ajustamento de Hoaglin, visualiza-se o não ajustamento dos dados à distribuição de Poisson e conjectura-se um possível ajustamento a uma mistura de distribuições de Poisson′s, tema que fica em aberto para posteriores estudos.

Capítulo 5

Teoremas Limite no Ensino da Matemática

5.1 Introdução

Nos actuais programas da disciplina de Matemática no Ensino Secundário é solicitado aos alunos que justifiquem processos de resolução, que encadeiem raciocínios, que confirmem conjecturas, demonstrem fórmulas e alguns teoremas.

Neste sentido, toda a actividade que envolva investigação e criação de conjecturas, revela-se de particular interesse, constituindo um modo privilegiado para reforçar uma introdução ao método científico.

Com um primeiro propósito de visualizar o comportamento da distribuição Binomial, a qual faz parte do actual programa de Matemática A do 12.º ano de escolaridade, no tema: Probabilidades e Análise Combinatória, são aqui apresentadas aplicações interactivas que relacionam a distribuição Binomial sobre vários pontos de vista. Pretende-se que o aluno, intuitivamente observe situações que conduzem a outras distribuições aproximadas para a distribuição Binomial e situações onde esta representa uma aproximação.

Estas aplicações têm o propósito de contribuir para a elaboração de novos suportes de apoio ao ensino das Probabilidades para os estudantes do Ensino Secundário, pois espera-se que toda e qualquer interactividade com carácter lúdico permita facilitar a apreensão dos conhecimentos. Acredita-se que a utilização das novas tecnologias, duma forma planeada e sistemática, permita:

- o desenvolvimento de competências de trabalho em autonomia. "Se é verdade que ne-

nhuma tecnologia poderá jamais transformar a realidade do sistema educativo, as novas tecnologias trazem dentro de si uma nova possibilidade: a de poder confiar realmente a todos os alunos a responsabilidade das suas aprendizagens (Carrier, J.-P., 1998)";

- uma prática de confrontação e verificação dos conhecimentos.

5.2 Tema: Probabilidades e Análise Combinatória no programa de Matemática

Há que contextualizar o ensino do tema: Probabilidades e Análise Combinatória no programa de Matemática A, 12.º ano, actualmente em vigor.

A disciplina de Matemática A surge no currículo, dos Cursos Gerais de Ciências Naturais, Ciências e Tecnologias e Ciências Socioeconómicas, como uma disciplina trienal da componente de Formação Específica, à qual é atribuída uma carga horária semanal de 4h 30min, dividida por aulas de 90 minutos, ao longo de 33 semanas lectivas.

Ao longo dos três anos do Ensino Secundário, os estudantes abordam as seguintes áreas: Números e Geometria, Funções Reais e Análise Infinitesimal; Estatística e das Probabilidades.

A abordagem da Estatística e das Probabilidades completa as aprendizagens básicas com algumas novas noções e ferramentas que não podiam ser compreendidas no Ensino Básico. As Probabilidades fornecem conceitos e métodos para estudar casos de incerteza e para interpretar previsões baseadas na incerteza. Este estudo, que pode ser em grande parte experimental, fornece uma base conceptual que capacita para interpretar, de forma crítica, toda a comunicação que utiliza a linguagem das probabilidades, bem como a linguagem estatística.

Cada vez mais para atingir os objectivos e competências gerais do programa dever-se-à recorrer às novas tecnologias. Não se trata de substituir o cálculo de papel e lápis pelo apoio da tecnologia, mas sim combinar adequadamente os diferentes processos de cálculo, sem esquecer o cálculo mental.

Na expressão feliz de Miguel de Guzmán, os estudantes devem ser preparados para um "diálogo inteligente com as ferramentas que já existem". O uso da tecnologia também facilita uma participação activa do aluno na sua aprendizagem como já era preconizado por Sebastião e Silva, quando escrevia que "haveria muitíssimo a lucrar em que o ensino ... fosse ... tanto quanto possível laboratorial, isto é, baseado no uso de computadores".

5.3 Aplicação Interactiva

Elaboraram-se aplicações interactivas que permitem visualizar a influência dos parâmetros nas representações gráficas das fmp′s das distribuições discretas: Bernoulli, Binomial, Poisson e Hipergeométrica e, que sugerem também aproximações entre elas e entre a função densidade de probabilidade da distribuição Normal.

É importante, ao nível do Ensino Secundário, incentivar a análise do comportamento de funções reais. Assim, propõe-se a análise das fmp′s daquelas distribuições discretas com a variação dos seus parâmetros. As aplicações desenvolvidas e apresentadas em D:\Componente Ensino\Atalho para Iniciar.pps, permitem então aos estudantes, para cada caso, observar a variação dos respectivos parâmetros proporcionando, de modo simples, uma correcta visualização do que acontece às funções.

Outro objectivo é estimular a criação de conjecturas sobre possíveis relações limite envolvendo distribuições discretas (Binomial, Poisson e Hipergeométrica) e também uma distribuição contínua muita útil em Estatística (a distribuição Normal). Pretende-se levar os estudantes a conjecturar, de modo puramente intuitivo, a convergência da distribuição Binomial à Normal, da distribuição Hipergeométrica à Binomial e desta à distribuição de Poisson.

Para atingir este objectivo, criaram-se três aplicações em *Java* envolvendo, propositadamente, em cada aplicação, as distribuições de especial interesse. Concretamente, uma aplicação elabora a representação gráfica das fmp′s das distribuições Binomial, $B(n,p)$ e Hipergeométrica, $H(N, M = p \times N, n)$; outra aplicação elabora a representação das fmp′s das distribuições Binomial, $B(n,p)$ e de Poisson, $P(\lambda = np)$; e uma terceira aplicação representa a fmp′s da distribuição Binomial $B(n,p)$ e a função densidade da distribuição Normal com média np e variância $np(1-p)$. Para cada um dos três casos pretende-se que os estudantes possam definir regras sobre os parâmetros das distribuições acerca das quais a aplicabilidade da aproximação é (visivelmente) válida.

Quando se recorrem a distribuições assimptóticas, como é o caso das três aplicações implementadas, levanta-se naturalmente o problema de averiguar a partir de que valores de n finito são desprezáveis os erros cometidos nas aproximações usadas. A resposta a esta questão, não é simples, e depende, logicamente, das estatísticas consideradas.

Aplicação Interactiva: Binomial → Normal

Sabe-se que a soma de v.a.′s normais independentes é ainda uma variável normal; sabe-se também que a distribuição aproximada da soma de n variáveis aleatórias independentes e identicamente distribuídas, desde que os momentos da sua distribuição verifiquem certas condições é também normal. É isso que diz o TLC, de que o teorema de De Moivre-Laplace é caso particular.

Formalmente:

Teorema do limite central

Dada a sucessão de variáveis aleatórias independentes e identicamente distribuídas $X_1, X_2, \cdots$ $\cdots$, $X_n, \cdots$, com valor médio μ e variância σ^2 (finita), então, quando $n \to \infty$, a função de distribuição da variável aleatória $Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$ tende para a função de distribuição da $N(0,1)$.

Demonstração. ver [39].

Teorema de De Moivre-Laplace

Se X é uma variável aleatória com distribuição Binomial, $X \sim B(n,p)$ com valor médio $\mu = np$ e variância $\sigma^2 = np(1-p)$, então, quando $n \to \infty$, a função de distribuição da variável aleatória $\frac{X-np}{\sqrt{np(1-p)}}$ tende para a função de distribuição da $N(0,1)$.

Demonstração. Consequência imediata do TLC quando os X_i, $i = 1, 2, \cdots$ seguem a distribuição $B(p)$.

Segundo alguns autores ([41]) a utilização da aproximação estabelecida neste dois teoremas limite pode resumir-se da seguinte maneira:

- se $n \leq 20$, deve utilizar-se directamente a distribuição binomial, o que permite o cálculo exacto das probabilidades;
- se $n > 20$, deve utilizar-se o cálculo aproximado das probabilidades, assim deve utilizar-se a aproximação da Normal à Binomial, considerando a correcção de continuidade, se $0.1 < p < 0.9$.

Espera-se que com a aplicação "Teorema Limite I - Convergência da Binomial à Normal" os utilizadores concluam estas regras práticas.

Aplicação Interactiva: Binomial → Poisson

A distribuição Binomial é utilizada quando se contabiliza o número de sucessos em n provas independentes. Porém, em algumas aplicações, o número de provas a considerar pode ser muito elevado e a probabilidade de sucesso muito pequena, tornando-se o cálculo de probabilidades muito moroso. Nestes casos, pode utilizar-se a distribuição de Poisson.
Formalmente:

Teorema de aproximação da Poisson à Binomial

A fmp da distribuição Binomial de parâmetros n e p, converge para a fmp da distribuição de Poisson de parâmetro λ do seguinte modo: $n \to \infty$, $p \to 0$ e $np \to \lambda$.

Demonstração. ver [39].

Quando n é "grande" e p é "pequeno", com np fixo, a fmp da distribuição Binomial pode ser aproximada pela fmp da distribuição de Poisson. Este resultado é muitas vezes conhecido pela designação de lei dos acontecimentos raros.
As regras práticas que se espera que os alunos formulem com a utilização da aplicação "Teorema Limite II - Convergência da Binomial à Poisson" podem resumir-se segundo alguns autores ([41]), da seguinte maneira:

- se $n \leq 20$, utiliza-se directamente a distribuição binomial, o que permite o cálculo exacto das probabilidades;
- se $n > 20$, deve utilizar-se o cálculo aproximado das probabilidades atendendo aos seguintes casos:
  - se $p \leq 0.1$, deve utilizar-se a aproximação da Poisson à Binomial;
  - se $p \geq 0.9$, também deve utilizar-se a aproximação da Poisson à Binomial, considerando o respectivo acontecimento complementar.

Aplicação Interactiva: Hipergeométrica → Binomial

As distribuições Binomial e Hipergeométrica são modelos teóricos adequados para estudar as propriedades de esquemas probabilísticos do seguinte tipo: de uma urna contendo N bolas, das quais M são do tipo "sucesso" e as restantes $N - M$ são do tipo "insucesso", considera-se a variável definida pelo número de bolas do tipo "sucesso" obtidas em n extracção de bolas. Conforme as extracções são feitas, com ou sem reposição, assim resultam os esquemas Binomial ou Hipergeométrico, respectivamente. A diferença fundamental entre os dois esquemas

é a seguinte: no primeiro, as extracções são independentes (as n provas de Bernoulli são independentes), dando lugar à distribuição Binomial; no segundo esquema, as provas não são independentes, uma vez que a probabilidade de extrair uma bola tipo "sucesso" na 1.ª prova é dada por M/N, enquanto que a probabilidade de sair uma bola tipo "sucesso" na 2.ª prova vai depender do resultado da primeira prova (se se extraiu uma bola tipo "insucesso", a probabilidade será dada por $M/(N-1)$; se a bola extraída tiver sido tipo "sucesso", a probabilidade será dada por $(M-1)/(N-1)$).

Formalmente:

Teorema de aproximação da Hipergeométrica à Binomial

A fmp da distribuição Hipergeométrica de parâmetros N, M e n, converge para a fmp da distribuição de Binomial de parâmetros n e p quando $N \to \infty$, $M \to \infty$ tal que $\frac{M}{N} \to p$.

Demonstração. ver [39].

Com a aplicação construída "Teorema Limite III - Convergência da Hipergeométrica à Binomial" espera-se que os estudantes cheguem à conjectura de que as distribuições se aproximam quando n é muito pequeno e N é muito grande, de tal modo que M/N se mantém constante (logo, necessariamente, $M \to \infty$).

É de referir que, em termos práticos, quando $n < N/10$ a distribuição Binomial fornece já uma aproximação satisfatória da distribuição Hipergeométrica.

Capítulo 6

Conclusão

Esta dissertação constitui uma abordagem no contexto dos codões, no estudo dos genomas, usando métodos estatísticos.

Com este capítulo conclui-se assim esta dissertação sumariando do ponto de vista mais biológico, os resultados obtidos ao longo do Capítulo 4, onde foram aplicadas várias metodologias estatísticas no estudo do genoma de várias espécies pertencentes aos três domínios da Vida.

O objectivo inicial da análise estatística realizada aos dados genómicos consistia na análise de propriedades associadas aos pares de codões iguais e codões raros para diversas espécies nos três domínios.

Até ao presente trabalho, os mapas de contextos de pares de codões de qualquer espécie tinham sido construídos definindo a significância dos codões segundo o critério de Bonferroni. Aqui outros procedimentos, associados a testes simultâneos, foram aplicados com o objectivo de confrontar a informação dos novos mapas com os até agora considerados. Os estudos apresentados na secção 4.2 mostram diferentes mapas para algumas espécies e levanta a questão se um critério baseado no controlo da taxa de falsos positivos, como é o de Bonferroni, conduz a uma perca relevante de informação que pode ser transmitida quando se utilizam critérios para controlar a taxa de falsas descobertas em testes simultâneos como é exemplo o procedimento de Storey.

Relativamente aos pares de codões iguais observou-se que é notoriamente significativa a presença de pares de codões coloridos em qualquer domínio da Vida, com especial incidência para os pares preferidos. Além disso, os resultados sumariados nas Tabelas 4.9, 4.10 e 4.11, destacam nas primeiras linhas os pares de codões iguais que são altamente problemáticos em cada domínio, quer por serem preferidos (*green*) ou preteridos (*red*).

Mais, da análise realizada constata-se que o genoma dá efectivamente preferência a pares de codões iguais quer em termos do peso da sua distribuição de frequência e quer face à independência. Por outras palavras, observa-se um maior número de pares de codões iguais do que aquele que seria esperado: *i)* se estivessem distribuídos uniformemente pelo genoma; e *ii)* se não existisse associação entre codões justapostos.

Relativamente aos codões raros não existem muitos estudos sobre eles, não se sabendo a sua importância no código genético. No presente trabalho conclui-se pelo não ajustamento da frequência de codões raros por gene a um modelo de Poisson, conjecturando-se um eventual ajuste a misturas de Poisson.

Os resultados obtidos, as ideias e direcções deixadas em aberto constituem uma boa base para futuras investigações.

Apêndices

Apêndice A

GENSCAN00000022707 cdna:Genscan chromosome:NCBI35:1:17480892:17509733:-1 - Espécie *H.sapiens*
AUGGCCCCGCCCCCACAGCUCCCUCUGCGCCCCUCCCCCCCCGCCCCGGGAAAGGGGACC
UCGAAACUUGAAGGGUCAAAGUGCAAAGGGCAGCUUUUGAUUUUUGGGGCAACCAACUGG
GACUUGAUUGGUCGAAAAGAAGUGCCUAAACAGCAAGCUGCUUACCGCAAUCUCGGUCAG
AAUUUGUGGGGGCCCCACAGAUAUGGGUGCCUGGCGGGGGUCCGGGUGCGGACAGUGGUC
UCGGGCUCGUGUGCUGCACACAGCCUCCUCAUCACCACGGAAGGGAAGCUGUGGAGCUGG
GGUCGAAAUGAGAAGGGGCAGCUGGGACAUGGUGACACCAAGAGAGUAGAAGCCCCUAGA
CUCAUCGAGGGUCUUAGCCACGAAGUGAUUGUGUCUGCAGCAUGUGGGCGGAACCACACC
UUGGCCUUGACGGAAACGGGCUCCGUGUUUGCGUUUGGGGAAAACAAGAUGGGGCAGCUG
GGCCUUGGCAACCAGACAGACGCUGUUCCCAGCCCCGCGCAGAUAAUGUACAACGGCCAG
CCAAUUACCAAAAUGGCCUGUGGGGCUGAAUUCAGUAUGAUAAUGGACUGCAAAGGAAAC
CUCUAUUCCUUUGGGUGCCCUGAAUAUGGUCAGCUGGGACACAACUCAGAUGGGAAGUUC
AUCGCCCGGGCACAGCGGAUAGAGUACGACUGUGAACUAGUUCCCCGGCGAGUGGCCAUC
UUCAUUGAGAAGACGAAAGAUGGACAGAUUCUGCCUGUACCAAACGUGGUUGUACGAGAC
GUGGCCUGUGGCGCUAACCACACGCUGGUCCUGGACUCCCAGAAGCGAGUCUUCUCCUGG
GGCUUUGGUGGCUAUGGCCGGCUGGGCCACGCAGAGCAGAAGGAUGAGAUGGUCCCCCGC
CUGGUGAAGCUGUUUGACUUCCCUGGGCGUGGGGCUUCCCAGAUCUAUGCUGGUUACACC
UGCUCCUUUGCUGUCAGUGAAGUGGGUGGUCUGUUUUUCUGGGGGGCCACCAACACCUCC
CGUGAAUCUACCAUGUACCCAAAAGCAGUGCAGGACCUCUGCGGCUGGAGAAUCCGGAGC
CUGGCUUGUGGGAAGAGCAGCAUCAUUGUGGCCGCCGAUGAGAGCACCAUCAGCUGGGGU
CCGUCACCGACCUUUGGGGAACUGGGCUACGGGGACCACAAGCCCAAGUCUUCCACUGCA
GCCCAGGAGGUAAAGACUCUGGAUGGCAUUUUCUCAGAGCAGGUCGCCAUGGGCUACUCA
CACUCCUUGGUGAUAGCAAGAGAUGAAAGUGAGACUGAGAAAGAGAAGAUCAAGAAACUG
CCAGAAUACAACCCCCGAACCCUCUGA

Apêndice B

Espécies do Domínio dos ARCHAEA consideradas para estudo na presente dissertação (18): Archaeoglobus fulgidus, Haloarcula marismortui, Halobacterium sp, Methanobacterium thermoautotrophicum, Methanococcus jannaschii, Methanococcus maripaludis, Methanosarcina acetivorans, Methanosarcina mazei, Picrophilus torridus, Pyrobaculum aerophilum, Pyrococcus abyssi, Pyrococcus furiosus, Pyrococcus horikoshii, Sulfolobus solfataricus, Sulfolobus tokodaii, Thermococcus kodakaraensis, Thermoplasma acidophilum e Thermoplasma volcanium.

Espécies do Domínio dos EUKARYA consideradas para estudo na presente dissertação (20): Arabidopsis thaliana, Candida albicans, Encephalitozoon cuniculi, Eremothecium gossypii, Plasmodium falciparum, Saccharomyces cerevisiae, Schizosaccharomyces pombe,Saccharomyces mikatae, Apis mellifera, Caenorhabditis elegans, Canis familiaris, Danio rerio, Takifugu rubripes, Gallus gallus, Homo sapiens, Mus musculus, Pan troglodytes, Rattus norvegicus,Tetraodon nigroviridis, Xenopus tropicalis.

Espécies do Domínio dos BACTERIA consideradas para estudo na presente dissertação (81): Acinetobacter sp, Agrobacterium tumefaciens, Aquifex aeolicus, Azoarcus sp, Bacillus cereus, Bacteroides fragilis, Bartonella henselae, Bdellovibrio bacteriovorus, Bordetella bronchiseptica, Bradyrhizobium japonicum, Brucella suis, Burkholderia pseudomallei, Campylobacter jejuni, Caulobacter crescentus, Chlorobium tepidum, Chromobacterium violaceum, Clostridium perfringens, Corynebacterium glutamicum, Coxiella burnetii, Dehalococcoides ethenogenes, Deinococcus radiodurans, Desulfotalea psychrophila, Desulfovibrio vulgaris, Enterococcus faecalis, Erwinia carotovora, Escherichia coli, Francisella tularensis, Fusobacterium nucleatum, Geobacillus kaustophilus, Geobacter sulfurreducens, Gloeobacter violaceus, Gluconobacter oxydans, Haemophilus ducreyi, Helicobacter hepaticus, Idiomarina loihiensis, Lactobacillus plantarum, Lactococcus lactis, Leifsonia xyli, Leptospira interrogans, Listeria innocua, Mannheimia succiniciproducens, Mesorhizobium loti, Methylococcus capsulatus, Mycobacterium bovis, Neisseria meningitidis, Nitrosomonas europaea, Nocardia farcinica, Oceanobacillus iheyensis, Parachlamydia, Pasteurella multocida, Photobacterium profundum, Photorhab-

dus luminescens, Pirellula sp., Porphyromonas gingivalis, Prochlorococcus marinus, Propionibacterium acnes, Pseudomonas syringae, Ralstonia solanacearum, Rhodopseudomonas palustris, Salmonella typhi, Shewanella oneidensis, Shigella flexneri, Silicibacter pomeroyi, Sinorhizobium meliloti, Staphylococcus aureus, Streptococcus, Streptomyces avermitilis, Symbiobacterium thermophilum, Synechococcus elongatus, Synechocystis, Thermoanaerobacter tengcongensis, Thermosynechococcus elongatus, Thermotoga maritima, Thermus thermophilus, Treponema denticola, Vibrio parahaemlyticus, Wolinella succinogenes, Xanthomonas campestris, Xylella fastidiosa, Yersinia pestis, Zymomonas mobilis.

Apêndice C

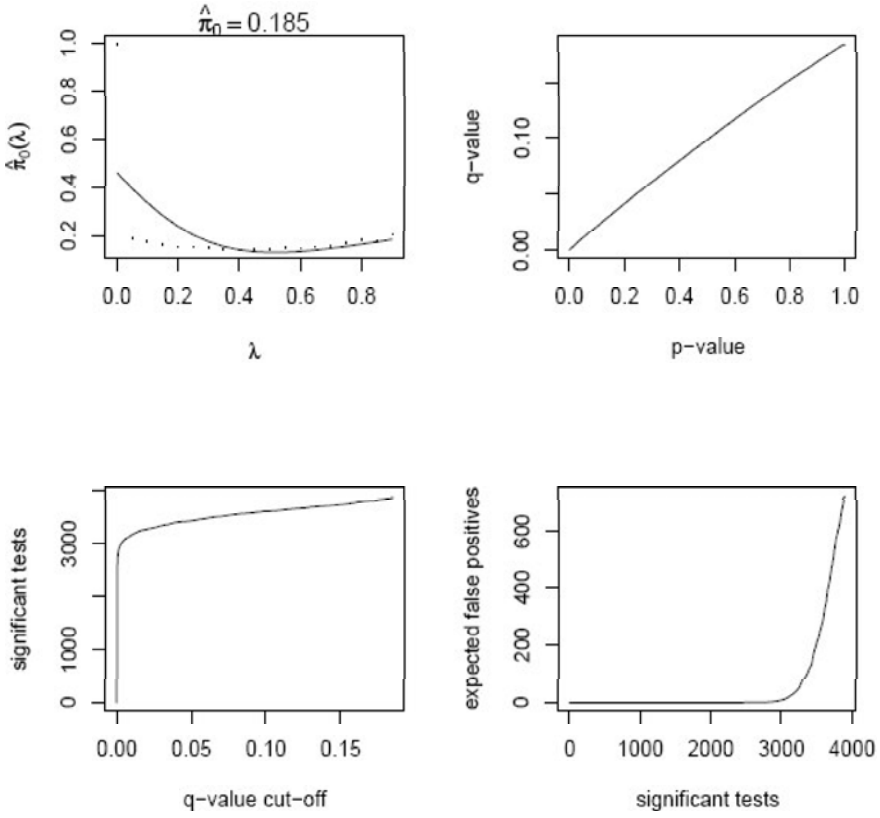


Figura 1: Resultados do conjunto de dados de pares de codões da espécie *Thermoplasma acidophilum*. (a) λ *vs* $\widehat{\pi}_0(\lambda)$. (b) $q-value$ *vs* o respectivo valor $p-value$. (c) número de testes significantes *vs* $q-value$ cut-off. (d) número esperado de falsos positivos *vs* número testes significantes.

H_i	Par codão	STAR	p-value	IND	BF	B-H	ST
1	AGA - UCC	82.982	0	S	S	S	S
2	GAC - AGG	64.646	0	S	S	S	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1078	CUC - ACG	4.719	0.000002	S	S	S	S
1079	UGC - GAC	4.684	0.000002	S	N	S	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1303	UUG - CGU	2.94	0.003261	S	N	S	S
1304	CGG - UGU	2.933	0.003282	S	N	N	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1365	AUA - AGC	2.582	0.009822	S	N	N	S
1366	UUC - CGA	2.574	0.010053	N	N	N	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1398	AUG - CCA	2.34	0.019283	N	N	N	S
1399	CGC - GUC	2.336	0.019491	N	N	N	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2206	ACA - GAG	-2.159	0.299270	N	N	N	N
2207	GUG - UGU	-2.162	0.298804	N	N	N	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2283	AUG - CUC	-2.575	0.010024	N	N	N	S
2284	AGA - CGA	-2.579	0.009908	S	N	N	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2311	AUA - UGC	-2.742	0.006410	S	N	N	S
2312	UAC - CAC	-2.756	0.005851	S	N	S	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2653	CGG - GGU	-4.698	0.000002	S	N	S	S
2654	GUG - UUG	-4.705	0.000002	S	S	S	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3904	GAU - AGG	-39.584	0	S	S	S	S

Tabela 1: Pares de codões preteridos/preferidos da espécie *Thermoplasma acidophilum*. Quatro procedimentos (Individual(IND), Bonferroni (BF), Bejamini-Hochberg (B-H) e Storey (ST)) para controlo das false positive rates correspondentes (FPR, FWER, FDR e pFDR) com $\alpha = 0.01$ são aplicados para encontrar pares significantes. S e N representam pares significantes e pares não significantes.

Procedimento	#pares preferidos	#pares preteridos
Teste Individual	1365	1621
Bonferroni	1078	1251
Bejamini-Hochberg	1303	1593
Storey	1398	1698

Tabela 2: Resumo dos resultados dos testes simultâneos - da espécie *Thermoplasma acidophilum*.

	Storey	B-H IND	BF
Y	56	55	52
$\widehat{p}$	0.918	0.902	0.852
I_N	0.849	0.827	0.763
	0.987	0.976	0.941
I_S	0.822	0.802	0.743
	0.964	0.954	0.92
I_{BS}	0.812	0.791	0.733
	0.969	0.959	0.926

Tabela 3: IC a 95% para p_j a probabilidade de um elemento da diagonal ser colorido para a espécie *Thermoplasma acidophilum*.

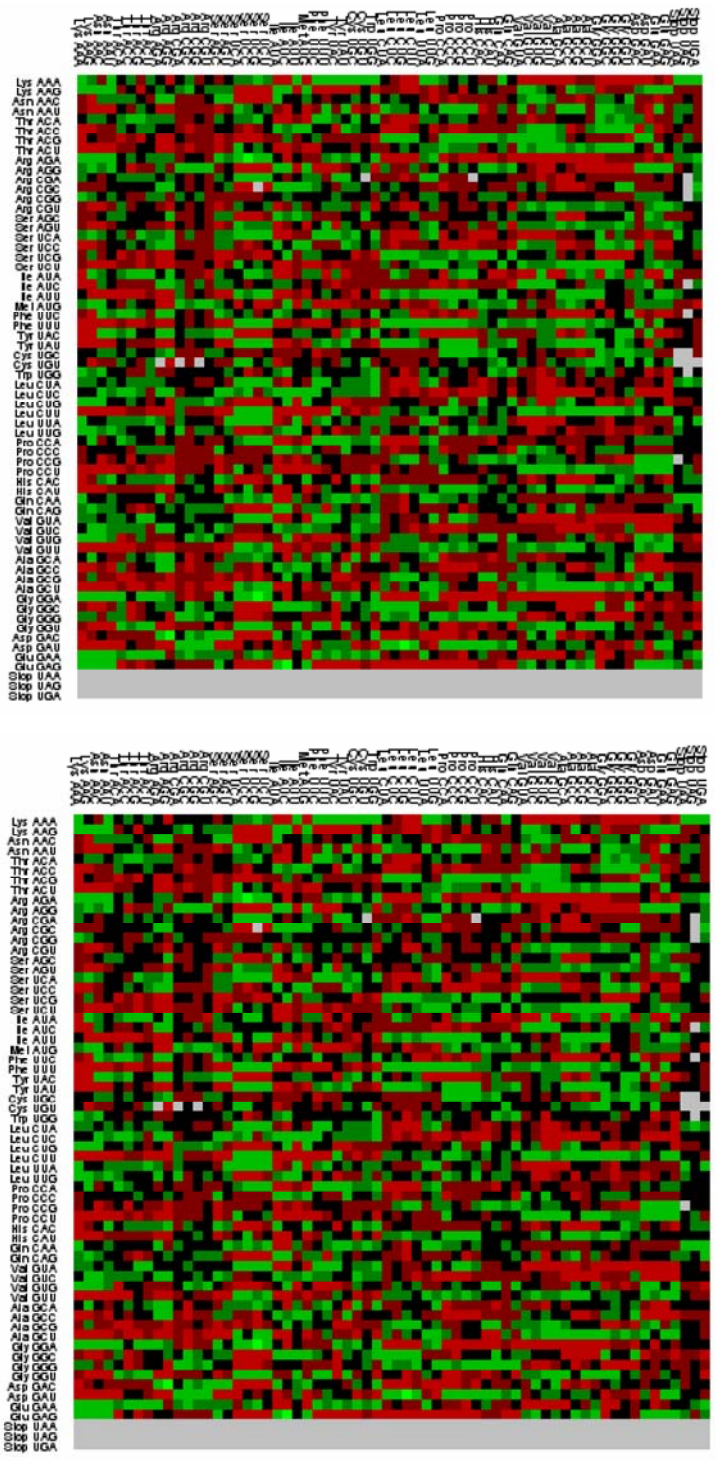


Figura 2: Mapa Contexto 3′ - Imagem correspondente à matriz dos resíduos da espécie *Thermoplasma acidophilum* c/ procedimentos de Storey e Bejamini-Hochberg, respectivamente.

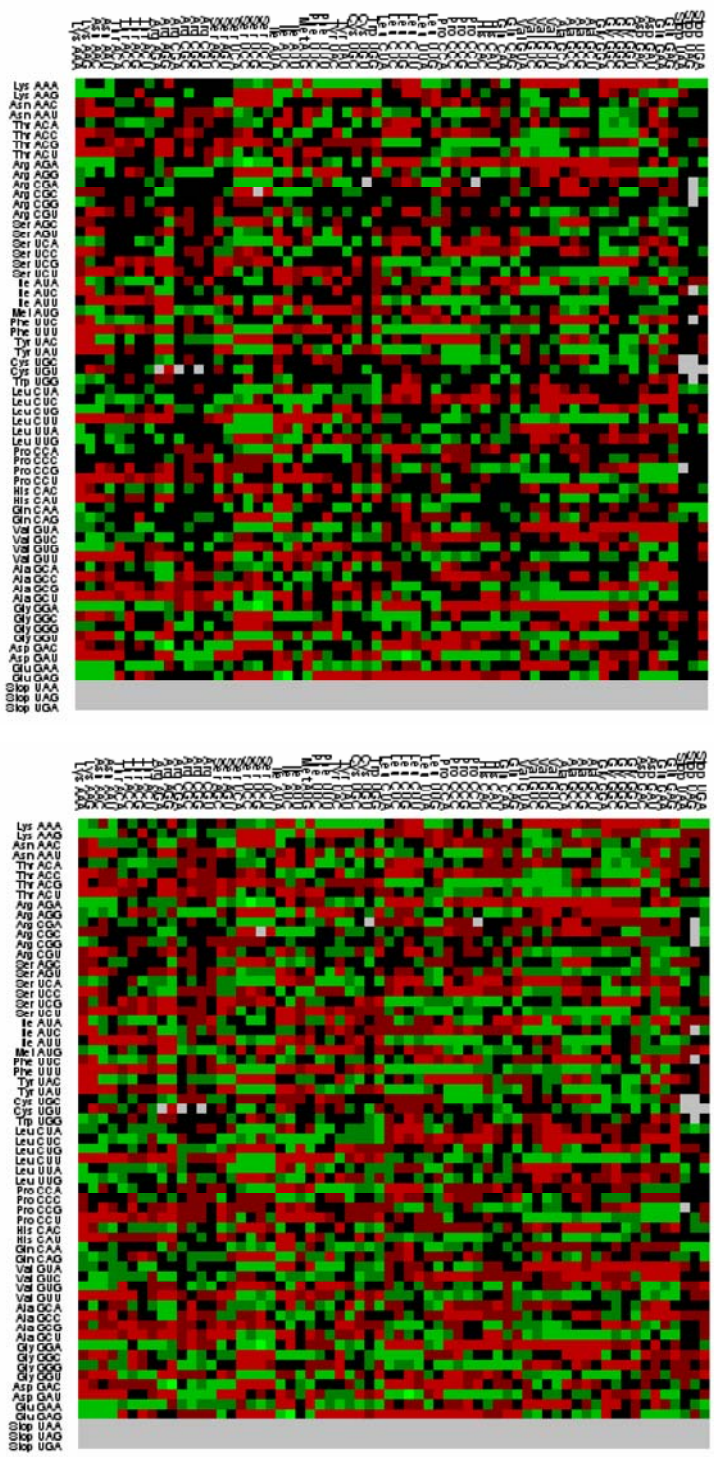


Figura 3: Mapa Contexto $3'$ - Imagem correspondente à matriz dos resíduos da espécie *Thermoplasma acidophilum* c/ procedimentos de Bonferroni e Individual, respectivamente.

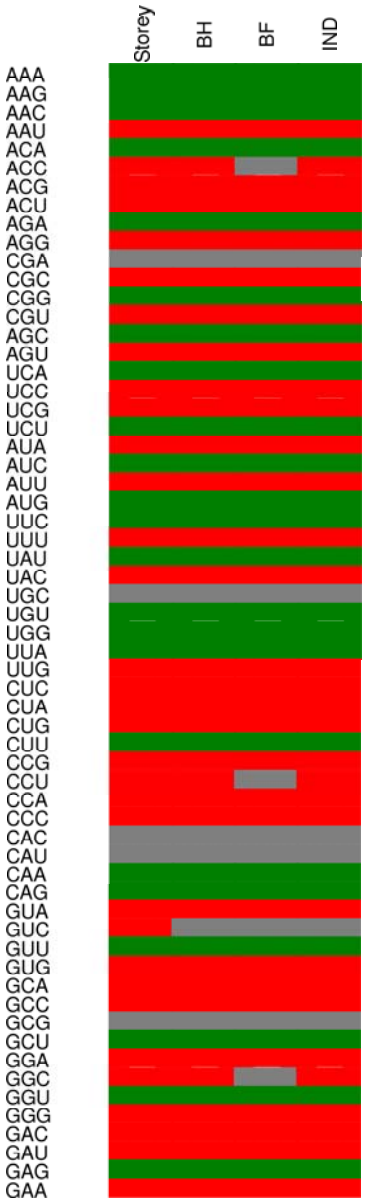


Figura 4: Diagonal principal da espécie *Thermoplasma acidophilum* segundo os procedimentos de ST, B-H, BF e IND, respectivamente.

	Storey	B-H IND	BF
Y_r	33	32	29
Y_g	23	23	23
Y_b	5	6	9
$\widehat{p}_r$	0.541	0.525	0.475
$\widehat{p}_g$	0.377	0.377	0.377
$\widehat{p}_b$	0.082	0.098	0.148
IC Gold			
L_r	0.385	0.368	0.319
U_r	0.697	0.681	0.632
L_g	0.225	0.225	0.225
U_g	0.529	0.529	0.529
L_b	-0.004	0.005	0.036
U_b	0.168	0.192	0.259
IC Q-Hurst			
L_r	0.388	0.373	0.328
U_r	0.686	0.672	0.627
L_g	0.243	0.243	0.243
U_g	0.533	0.533	0.533
L_b	0.029	0.038	0.068
U_b	0.21	0.23	0.29
IC Goodman			
L_r	0.391	0.376	0.331
U_r	0.684	0.669	0.624
L_g	0.245	0.245	0.245
U_g	0.53	0.53	0.53
L_b	0.03	0.039	0.07
U_b	0.201	0.227	0.286

Tabela 4: IC a 95% para $p_j = (p_r, p_g, p_b)$ a probabilidade de um elemento da diagonal ser $r/g/b$ para a espécie *Thermoplasma acidophilum*.

Apêndice D

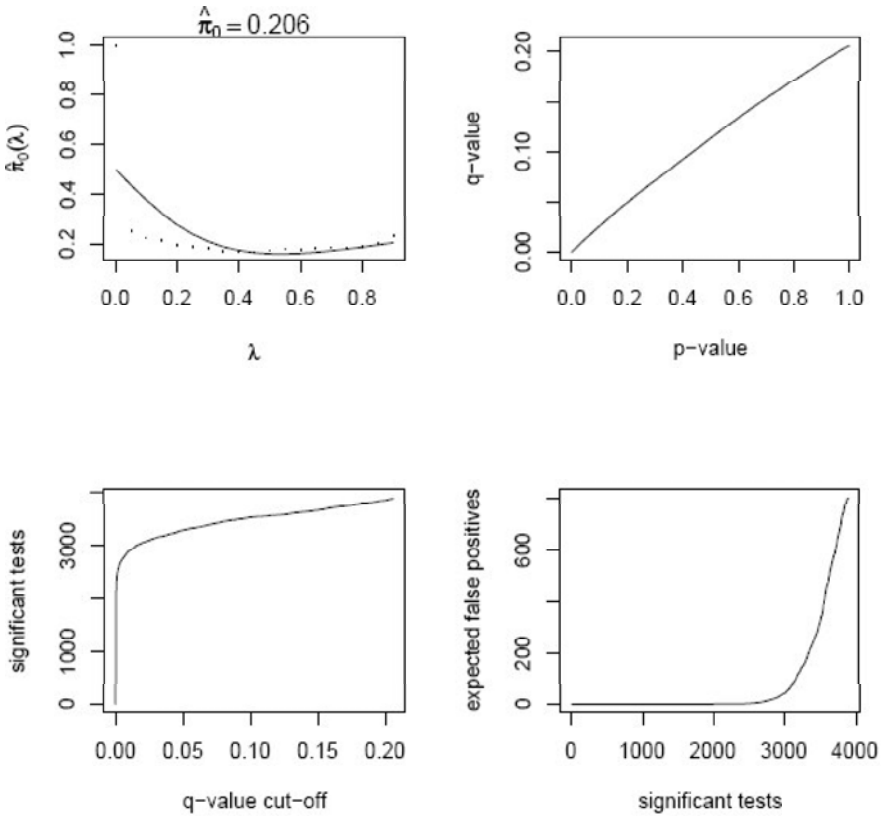


Figura 5: Resultados do conjunto de dados de pares de codões da espécie *Bacillus Cereus*. (a) λ *vs* $\widehat{\pi}_0(\lambda)$. (b) $q - value$ *vs* o respectivo valor $p - value$. (c) número de testes significantes *vs* $q - value$ cut-off. (d) número esperado de falsos positivos *vs* número testes significantes.

H_i	Par codão	STAR	p-value	IND	BF	B-H	ST
1	AAA - GAA	64.845	0	S	S	S	S
2	GUG - AAA	57.996	0	S	S	S	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
882	UUA - ACU	4.713	0.000002	S	S	S	S
883	AAA - UGG	4.695	0.000002	S	N	S	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1157	GUC - GCA	2.98	0.002882	S	N	S	S
1158	ACG - CUA	2.966	0.003017	S	N	N	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1226	ACG - CUG	2.579	0.009908	S	N	N	S
1227	GCG - ACA	2.574	0.010053	N	N	N	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1255	GCG - CUG	2.423	0.015392	N	N	N	S
1256	UUU - UAA	2.415	0.015735	N	N	N	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2353	GGU - GAU	-2.179	0.029331	N	N	N	N
2354	CCU - CGC	-2.189	0.028596	N	N	N	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2456	AGA - CUC	-2.573	0.010082	N	N	N	S
2457	GAU - CCC	-2.576	0.009995	S	N	N	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2500	CAG - AGA	-2.715	0.006627	S	N	N	S
2501	CAG - UGU	-2.728	0.006371	S	N	S	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2940	CAA - CGG	-4.689	0.000001	S	N	S	S
2941	CCA - AAG	-4.708	0.000001	S	S	S	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3904	GUU - AAA	-39.626	0	S	S	S	S

Tabela 5: Pares de codões preteridos/preferidos da espécie *Bacillus Cereus*. Quatro procedimentos (Individual(IND), Bonferroni (BF), Bejamini-Hochberg (B-H) e Storey (ST)) para controlo das false positive rates correspondentes (FPR, FWER, FDR e pFDR) com $\alpha = 0.01$ são aplicados para encontrar pares significantes. S e N representam pares significantes e pares não significantes.

Procedimento	#pares preferidos	#pares preteridos
Teste Individual	1226	1448
Bonferroni	882	964
Bejamini-Hochberg	1157	1404
Storey	1255	1551

Tabela 6: Resumo dos resultados dos testes simultâneos - da espécie *Bacillus Cereus*.

	Storey	BF	B-H IND
Y	53	49	52
$\widehat{p}$	0.869	0.803	0.852
I_N	0.784	0.704	0.763
	0.954	0.903	0.941
I_S	0.762	0.687	0.743
	0.932	0.884	0.92
I_{BS}	0.752	0.678	0.733
	0.938	0.89	0.926

Tabela 7: IC a 95% para p_j a probabilidade de um elemento da diagonal ser colorido para a espécie *Bacillus Cereus*.

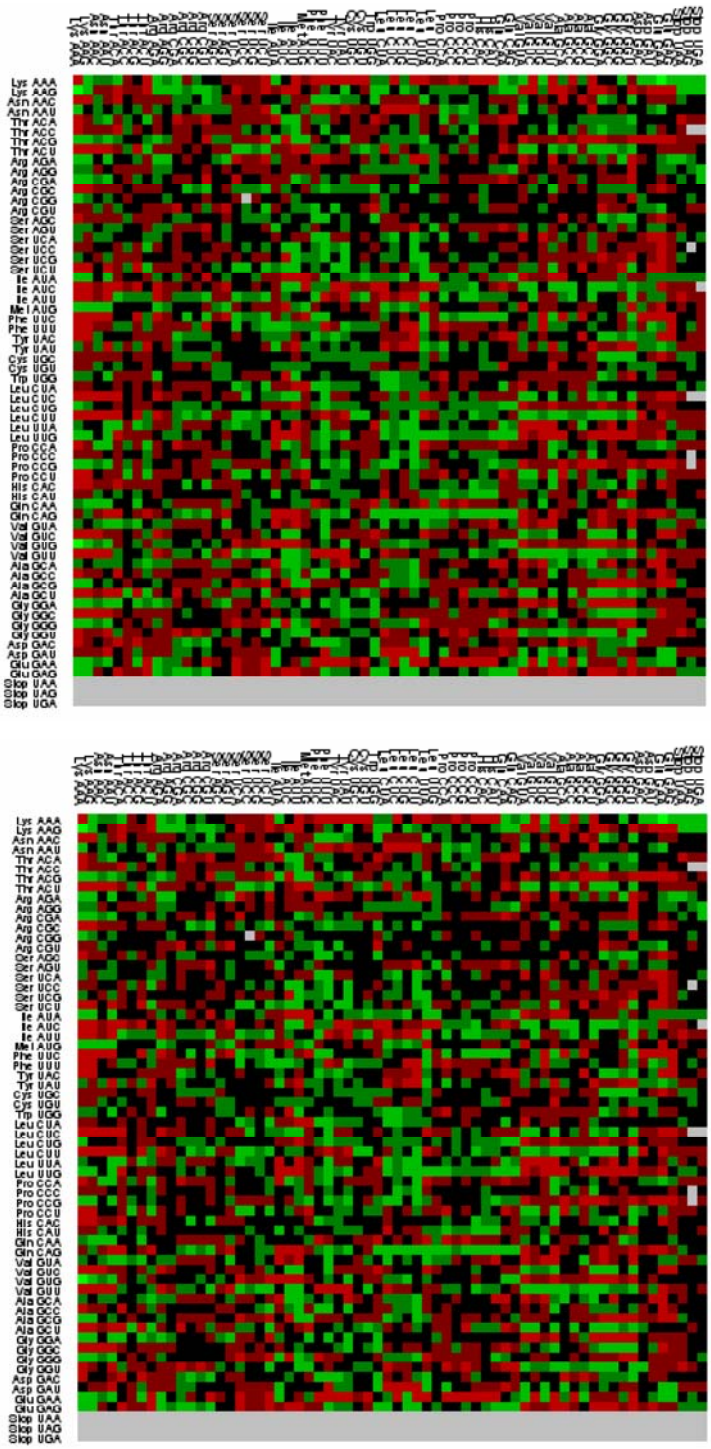


Figura 6: Mapa Contexto 3′ - Imagem correspondente à matriz dos resíduos da espécie *Bacillus Cereus* c/ procedimentos de Storey e Bejamini-Hochberg, respectivamente.

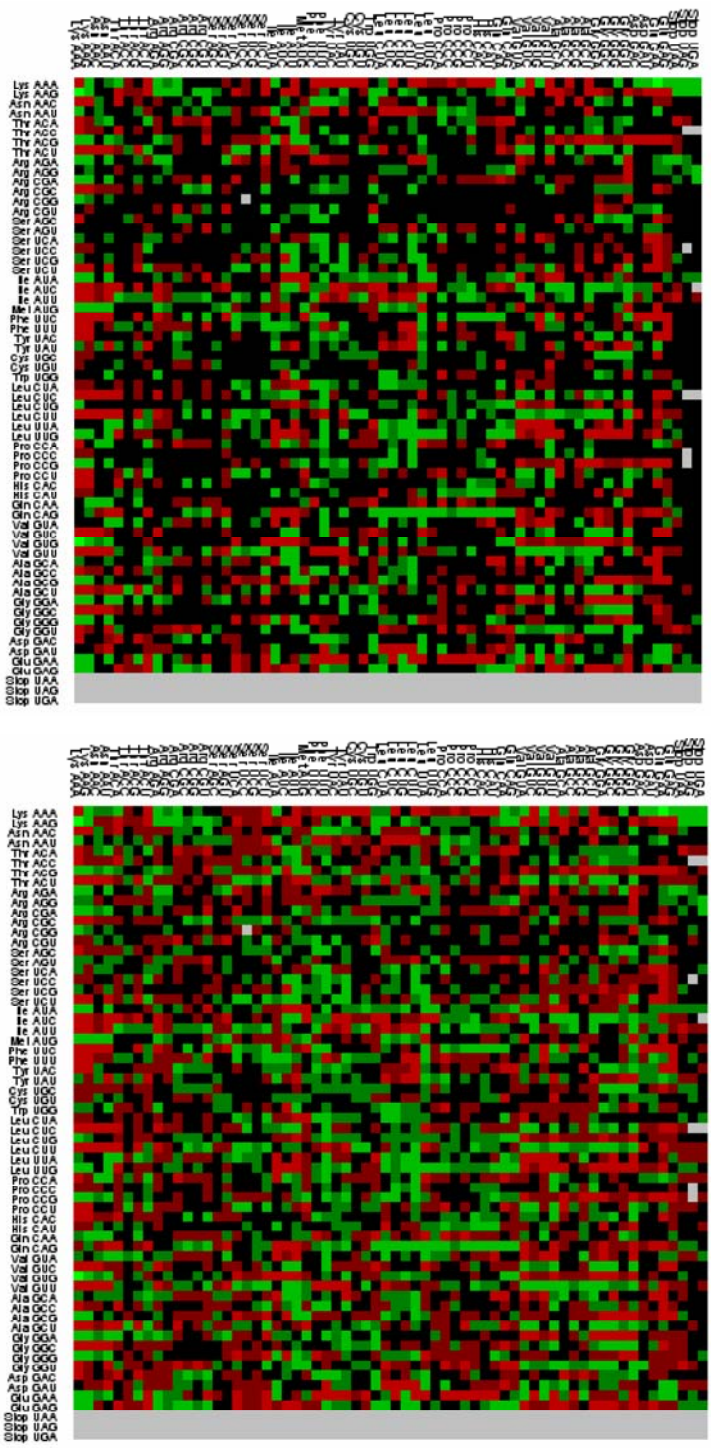


Figura 7: Mapa Contexto 3′ - Imagem correspondente à matriz dos resíduos da espécie *Bacillus Cereus* c/ procedimentos de Bonferroni e Individual, respectivamente.

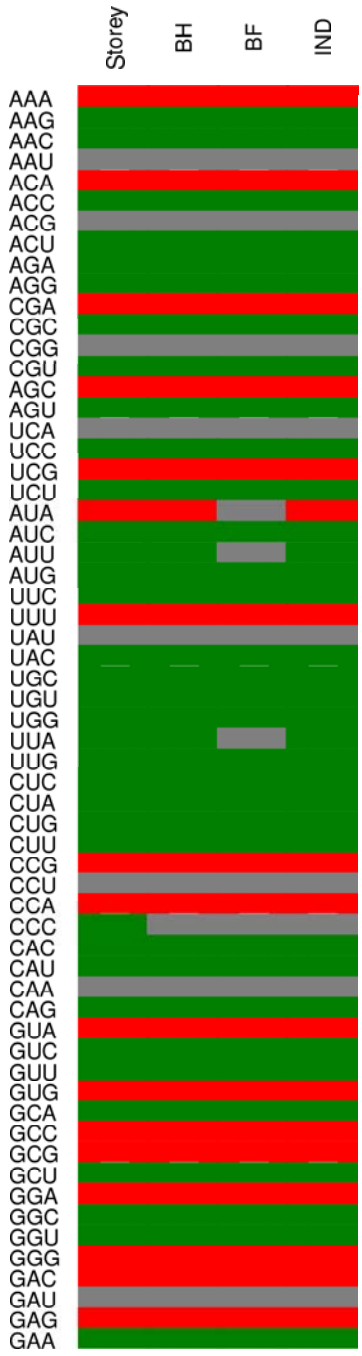


Figura 8: Diagonal principal da espécie *Bacillus Cereus* segundo os procedimentos de ST, B-H, BF e IND, respectivamente.

	Storey	B-H IND	BF
Y_r	17	17	16
Y_g	36	35	33
Y_b	8	9	12
$\widehat{p}_r$	0.279	0.279	0.262
$\widehat{p}_g$	0.59	0.574	0.541
$\widehat{p}_b$	0.131	0.148	0.197
IC Gold			
L_r	0.138	0.138	0.124
U_r	0.419	0.419	0.400
L_g	0.436	0.419	0.384
U_g	0.744	0.729	0.697
L_b	0.025	0.036	0.072
U_b	0.237	0.259	0.321
IC Q-Hurst			
L_r	0.163	0.163	0.15
U_r	0.434	0.434	0.417
L_g	0.435	0.419	0.388
U_g	0.729	0.715	0.686
L_b	0.058	0.068	0.102
U_b	0.27	0.29	0.346
IC Goodman			
L_r	0.165	0.165	0.152
U_r	0.430	0.430	0.413
L_g	0.438	0.422	0.391
U_g	0.727	0.712	0.684
L_b	0.059	0.07	0.103
U_b	0.267	0.286	0.342

Tabela 8: IC a 95% para $p_j = (p_r, p_g, p_b)$ a probabilidade de um elemento da diagonal ser $r/g/b$ para a espécie *Bacillus Cereus*.

Bibliografia

[1] Afreixo, V. (2002). *Análise Estatística da Linguagem Genética*. Tese de Mestrado. Aveiro: Universidade de Aveiro.

[2] Agresti, A. (2002). *Categorical data analysis*. New York: John Wiley and Sons, Inc.

[3] Azevedo, C. (1994). *Biologia Celular*. LIDEL - edições técnicas.

[4] Bejamini, Y. and Hochberg, Y. (1995). *Controlling the false discovery rate: a pratical and powerful approach to multiple testing*. J. R. Stat. Soc.. Ser. B. **57**: 289-300.

[5] Bejamini, Y. and Yekutieli, D. (2001). *The control of the false discovery ratein multiple testing under dependency*. Annals os Statistics **29**: 1165-1188.

[6] Berkson, J. (1938). *Some difficulties of interpretation encountered in the application of de chi-square test*. J. Amer. Statist. Assoc.. **33**: 526-536.

[7] Blyth, C. (1986). *Approximate Binomial Confidence Limits*. J. Amer. Statist. Assoc.. **81**: 843-855.

[8] Blyth, C. and Hutchinson, D. (1960). *Tables of Neyman-Shortest Unbiased Confidence Intervals for the Binomial Parameter*. Biometrika. **47**: 381-391.

[9] Blyth, C. and Still, H. (1983). *Binomial Confidence Intervals*. J. Amer. Statist. Assoc.. **78**: 108-116.

[10] Brown, L. and Li, X. (2004). *Confidence intervals for two sample binomial distribution*. Journal of Statistical Planning and Inference.

[11] Carey, F. (1996). *Organic Chemistry*. McGraw-Hill.

[12] Casella, G. (1986). *Refining Binomial Confidence Intervals*. Canad. J. Statist.. **14**: 113-129.

[13] Clopper, C. and Pearson, E. (1934). *The Use of Confidence or Fiducial Limits Illustrated in the Case of Binomial.* Biometrika. **26**: 404-413.

[14] Cochran, W. (1952). *The Chi-squared Test of Goodness of Fit.* Ann. Math. Statist. **23**: 315-345.

[15] Cochran, W. (1954). *Some methods of streng thening the common χ^2 tests.* Biometrics. **10**: 417-451.

[16] Costa, B., Resende,L. e Rodrigues, E. (2005). *Espaço 12.* Edições ASA.

[17] Cressie, N. and Read, T. (1984). *Multinomial Goodness-of-Fit Tests.* J. Roy. Statist. Soc. (B) **46**: 440-464.

[18] Crow, E. (1956). *Confidence Limits for a Proportion.* Biometrika. **43**: 423-435.

[19] Cunha, L.. *Dossiers Didácticos - VII Probabilidades com Excel.* Projecto ALEA.

[20] Daintith, J.. *Dicionário Breve de Química.* Editorial Presença.

[21] Efron, B. (2004). *Large-scale simultaneous hypothesis testing: the choice of a null hypothesis.* J. Amer. Statist. Assoc.. **99**: 99-104.

[22] Fujino, Y. (1980). *Approximate Binomial Confidence Limits.* Biometrika. **67**: 677-681.

[23] Fujino, Y. and Okuno,T. (1984). *The Minimax Average Confidence Limits for a Binomial Probability One-Sided Case.* Rep. Statist. Appl. Res. Juse. **31**: 1-7.

[24] Gart, J. (1964). *The analysis of Poisson Regression with an Application in virology.* Biometrika. **51**: 517-521.

[25] Glosh, B. (1979). *A Comparison of Some Approximate Confidence Intervals for the Binomial Parameter.* J. Amer. Statist. Assoc.. **74**: 894-900.

[26] Gold, R. (1963). *Test Auxiliary to χ^2 Tests in a Markov Chain.* Ann. Stat. **34**: 56-74.

[27] Goodman, L. (1965). *On Simultaneous Confidence Intervals for Multinomial Proportions.* Technometrics. **7**: 247-254.

[28] Haberman, S. (1973). *The analysis of residuals in cross-classified tables.* Biometrics. **29**: 205-220.

[29] Hoaglin, D. (1980). *A Poissonness Plot.* The American Statistician. **34**: 146-149.

[30] Jorge, A., Alves, C., Fonseca, G. e Barbedo, J. (2005). *Infinito 12º*.Areal Editores.

[31] Karlin, S. and Taylor, H. (1975). *A first course in Stochastic Processes.* New York: Academic Press.

[32] Kim, S., Tsui, K. and Borodovsky, M. (2005). *Multiple Testing in Large-Scale Contingency Tables: Inferring Pair-Wise Amino Acid Patterns in β -Sheets.* Journal of Statistical Planning and Inference.

[33] Lancaster, M. (1949). *The derivation and partition os χ^2 in certain discrete distribuitions.* Biometrika. **36**: 117-129.

[34] McLachln, G., Do, K. and Ambroise, C. (2004). *Analyzing microarray gene expression data.* Edições Wiley Inter-Science.

[35] Mendes, A. e Marcelino, M. (2003). *Fundamentos de Programação em JAVA 2, segunda edição.* Edições FCA - Editora de Informática.

[36] Ministério da Educação, Departamento do Ensino Secundário. *Brochura de Probabilidades 12.º Ano.*

[37] Ministério da Educação. *Programa homologado da disciplina de Matemática A do 12º ano.*

[38] Moura, G., Pinheiro, M., Silva, R., Miranda, I., Afreixo, V., Dias, G., Freitas, A., Oliveira, J.L. and Santos, M. (2005). *Comparative context analysis of codon pairs on an ORFeome scale.* Methods Inf Med. **6**: R28.

[39] Murteira, B. (1990). *Probabilidades e Estatística. Volume 1 segunda edição revista.*McGraw-Hill.

[40] Murteira, B. (1990). *Probabilidades e Estatística. Volume 2 segunda edição revista.*McGraw-Hill.

[41] Murteira, B., Silva, C., Silva, J. e Pimenta, C. (2001). *Introdução à Estatística.*McGraw-Hill.

[42] Neves, M., Guerreiro, L. e Moura, A. (2005). *Matemática A, 12º ano.* Porto Editora.

[43] Newcombe, R. (1998). *Interval Estimation for the difference between independent proportions: comparison of eleven methods.* Statist. Med. **17**: 873-890.

[44] Nossal, G. (1985). *A Engenharia Genética.* Editorial Presença.

[45] Ord, J. (1967). *Graphical Methods for a Class of Discrete Distribuitions.* J. Roy. Stat. Soc.. Ser. A. **130**: 232-238.

[46] Pereira, A. (2004). *SPSS - Guia Prático de Utilização, quinta edição revista e aumentada.*Edições Sílabo.

[47] Pinheiro, M., Moura, G., Silva, R., Miranda, I., Afreixo, V., Dias, G., Freitas, A., Oliveira, J.L. and Santos M. (2006). *Statistical, Computational and Visualization methodologies to unveil gene primary structure features.* Methods Inf Med. **2**: 163-168.

[48] Quesenberry, C. and Hurst, D. (1964). *Large Sample Simultaneous Confidence Intervals for Multinomial Proportions.* Technometrics. **6**: 191-195.

[49] Read, T. and Cressie, N. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data.* New-York: Springer-Verlag.

[50] Santner, T. and Duffy, D. (1989). *The Statistical Analysis of Discrete Data.* Springer-Verlag.

[51] Shaffer, J. (1995). *Multiple hypothesis testing.* Annu. Rev. Psychol.. **46**: 561-584.

[52] Sheffe, H. (1959). *The analysis of variance.* New York: John Wiley and Sons, Inc.

[53] Silva,A., Gramaxo, F., Santos, M., Mesquita, A. e Baldaia, L. (2000). *Terra Universo de Vida - 11º ano, 1ª Parte - Biologia.* Porto Editora.

[54] Sterne, T. (1954). *Some Remarks on Confidence or Fiducial Limits.* Biometrika. **41**: 275-278.

[55] Storey, J. (2002). *A direct approach to false discovery rates.*University of Stantford. Journal of the Royal Statistical Society, Series B. **64**: 479-498.

[56] Storey, J. (2003). *The positive false discovery rate: a bayesian interpretation and the q-value.*University of Washington. Annals of Statistics. Vol. **31**: 2013-2035.

[57] Storey, J. and Dabney, A. (2004). *Bioconductor's qvalue package.* Department of Biostatistics. University of Washington.

[58] Storey, J., Taylor, J. and Siegmund, D. (2004). *Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach.* University of Washington. Journal of the Royal Statistical Society, Series B. **66**: 187-205.

[59] Storey, J. and Tibshirani,R. (2003). *Statistical significance for genomewide studies.* Department of Biostatistics. University of Washington.

[60] Valente, A., Santos, M., Moura, G., Oliveira, J. L., Pinheiro, M. and Afreixo, V. (2004). *Association Between genetic symbols.* Departamento de Matemática da Universidade de Aveiro.

[61] Vos, J. (1978). *Confidence Intervals for a Binomial Parameter.* ISO/TC 69/SC 2/N 165.

[62] Vos, J. (1979). *Average Confidence Intervals.* ISO/TC 69/SC 2/N 175.

[63] http://www.apm.pt.

[64] http://www.des.pt.

[65] http://www.nctm.org.

Os anexos desta tese apenas poderão ser consultados através do CD-ROM.

Por favor queira dirigir-se ao 4º piso da Biblioteca e solicitá-lo no balcão de atendimento.